

IN DEEP WATER Rising sea levels set to outpace growth rate of coral reefs PAGES 378 & 396



EVIDENCE SYNTHESIS

How analysis can inform policy PAGES 361 & 364

THE HUNT FOR MISSING MATTER

Reservoir of baryons located in intergalactic medium PAGES 375 & 406

CELLULAR

An umbrella to protect stem cells from ultraviolet rays PAGES 374 & 445

Vol. 558, No. 7710

THIS WEEK

EDITORIALS

BIAS Figures show women are still under-represented in *Nature*'s pages **p.344**

WORLD VIEW Tell the truth about how impact factor matters **p.345**



The void in opioid research

The National Institutes of Health's plans to tackle the opioid epidemic in the United States can treat only the symptoms, not the cause.

alls for urgent action on the opioid crisis in the United States have been coming thick and fast. So, too, have the possible solutions: Congress is currently considering 57 opioid-related bills. This comes after President Donald Trump declared the crisis a public-health emergency last year. No one would dispute that. In 2016, more than 53,000 people in the United States died from an opioid overdose — more than double the figure in 2010 — and the increasing use, misuse and abuse of heroin, fentanyl and other opiates, including prescription drugs, shows no signs of slowing.

Action is needed — but what? The problem is that nobody can agree on possible solutions. Indeed, some of the proposals passing before Congress this week have conflicting intentions on matters such as how best to implement addiction treatment. Only rigorous research and evidence can steer this debate and identify the most effective ways to intervene. Yet, so far, a series of White House commissions has done little but talk.

Congress did, at least, start to put real money towards the issue this year — a total of US\$4.6 billion, including an extra \$500 million for research at the US National Institutes of Health (NIH). And this month, NIH director Francis Collins and colleagues laid out their plans to spend this latter windfall (F. S. Collins *et al. J. Am. Med. Assoc.* http://doi.org/cq38; 2018).

The agency's initiative is called Helping to End Addiction Longterm (HEAL) and divides its research strategy into two prongs: improving treatments for addiction and overdose, and improving pain management. The plan could have great value, but unfortunately it includes some questionable priorities.

On the positive side, the NIH plans to spend nearly 20% of HEAL funding on an initiative to test public-health interventions — such as better screening for addiction — through partnerships with emergency departments, justice systems and other sectors. It will spend \$10 million on developing and improving therapies for babies who are born addicted to opioids, and about \$29 million on expanding and improving its network of clinical trials for various therapies. A little under half of the money will be carried forward to the next financial year, to be used for praiseworthy programmes including prevention research, precision medicine for pain and addiction, and non-pharmacological and integrated models of pain management.

All good. Yet a great deal of the money will go towards drug development — and that's a less essential investment. The NIH would be better served by determining how best to deploy existing treatments, instead of spending years on expensive efforts to develop new ones. Current overdose-reversal drugs, such as naloxone, work extremely well, although access in an overdose situation remains a problem. Non-addictive painkillers such as paracetamol and ibuprofen can, in certain combinations, be as effective as opioids for some kinds of pain, but there is great need for improved scientific understanding, particularly of chronic pain.

Meanwhile, some ethicists have criticized the NIH's agenda as overly friendly to the pharmaceutical industry. Many critics argue that the

industry had a major role in starting the epidemic in the first place, by promoting drugs such as OxyContin (oxycodone) as non-addictive. To its credit, the NIH has stepped back from its initial plan for HEAL, which involved a direct partnership with industry, combining public and private money to fund drug development. In 2017, the agency held a series of closed-door meetings with the Food and Drug Administration and dozens of pharmaceutical companies, including representatives from opioid manufacturers Purdue Pharma of Stamford, Connecticut, and Janssen, headquartered in Beerse, Belgium. Both companies are facing multiple lawsuits from US states for deceptive marketing and hiding reports of adverse events. (In an interview with *Nature* last month, Collins said that he had invited industry repre-

"The epidemic's roots are a complex tangle of social and political issues."

sentatives because "we may as well hear what the various companies had to offer in terms of ways to address this public-health crisis".)

The NIH reversed course this April, on the advice of an ethics committee that recommended the agency refrain from taking cash from industry partners. That advice came

soon after revelations that NIH-funded researchers and employees convening a study on whether alcohol could improve health had courted funding from the alcohol industry — a practice forbidden by NIH policy. On 15 June, the agency announced that it had terminated the study.

The NIH's revised plan for HEAL will fund opioid research exclusively with federal money, and will involve industry partners only in setting up a clinical-trial network for drug testing and a system for sharing biomarkers. The agency will not partner with companies involved in litigation related to the opioid crisis. However, any industry relationship still has potential for a conflict of interest or undue influence. Transparency over those relationships — and continued federal funding — will be key to avoiding that. Pharmaceutical companies, meanwhile, should do their part by participating fully in HEAL research and by sharing data openly.

Even at their best, the NIH's findings will be able only to alleviate the symptoms of the opioid epidemic — helping people who are already addicted. What they cannot do is tackle its roots, which are a complex tangle of social and political issues including economic disparities, lack of access to comprehensive health care and mental-health services, outdated policies banning evidence-based initiatives such as local safe-injection facilities, a proliferation of deadly synthetic drugs and poor prescribing practices by physicians.

Curing the opioid epidemic requires funding, new public-health initiatives and enforcement of policies that address these problems. Drug overdose is now the leading cause of death for under-50s in the United States. With just 4% of the world's population, the country accounts for around 27% of all global overdose deaths. No matter how many new drugs are developed, only evidence-based policy — and the political will to enforce it — can begin to prevent this modern tragedy from spiralling further out of control.

Bias revisited

Women continue to represent too small a proportion of this journal's authors and referees.

both in its goals and in its actions, *Nature*'s editorial team is trying to address the issue of equity in science. See, for example, an Editorial published earlier this month (*Nature* 558, 5; 2018) and a collection of content from across the Nature group of journals (see go.nature.com/2gjwkkn).

As a part of this effort, we have previously provided statistics and regular updates on the balance between male and female contributors to *Nature* content, both as authors and as referees. Consistently, these have shown the involvement of too few women when compared with estimates of the number of females present in research communities. (As one indicator, data from the United Nations Educational, Scientific and Cultural Organization show that the global average proportion of women in the science workforce is about 29%; see go.nature. com/2koxupq.)

Since we published our first report on this topic in a 2012 Editorial (*Nature* **491**, 495; 2012), the numbers show we have made some progress, but not enough and too slowly. A key element has been our attempt to counter unconscious bias, by getting senior staff and editors to ask themselves, 'Who are the outstanding women for this task?', before commissioning an author or a referee. We cannot claim that this important exercise happens on every occasion, but we have made substantial efforts.

So what do the latest statistics reveal? The sections of *Nature* that are directly commissioned by in-house editors are where we have most agency, and so have been most responsive to our efforts. In 2017, in our Comment, World View, Books & Arts and Obituary sections, 29% of our 255 authors were women. The proportion of women authors in Comment and World View in 2017 was 34% — an increase since the 19% recorded in 2012.

These articles are commissioned by a team (all female, as it happens) that (like many others) works hard to deliver on this agenda. They

report a noticeable tendency for senior women to decline invitations. As was detailed in our 2012 Editorial, there are many reasons why women researchers might have less time for such writing than have men. The team also finds that advisers and invitees, whatever their gender, often send all-male suggestions for alternative authors. We are countering this latter tendency by asking all those who suggest authors or referees to "bear diversity in mind".

"There are many reasons why women researchers might have less time for such writing than have men."

The News & Views section of *Nature* has considerably improved its position with commissions since we started our initiatives in 2012, when the proportion of women authors stood at 12%. But over the past 3 years, despite keeping up its efforts, that ratio has plateaued at about 26% female — 113 out of 442 authors in 2017.

In the 47 Review articles that we published in 2017, from a total of 217 authors, 42 of them — just over 19% — were women.

Our poorest outcome is in the refereeing of research papers. Counting only individuals whose gender we can attribute from their first names, the proportion of female referees has increased from 12% in 2011 to 16% in 2017.

When assigning gender, we used the algorithm from Gender API. We counted records for which the algorithm could not match the name with a gender, or returned an accuracy below 95%, as 'unknown gender'. These results are skewed because the algorithm has a hard time identifying gender in some languages, such as Chinese. We counted the referees for all submissions — if a referee reviewed three different manuscripts in a given year, we counted them as three, not

For authors, we counted the total number of corresponding authors in a similar fashion. Counting only authors with an assignable gender, the percentage of female corresponding authors has remained constant at 16% over time.

The editors of *Nature* and of all the Nature journals have, in recent months, been consolidating existing initiatives on diversity and inclusion, and setting up new ones, to become more systematic and creative about this. We will report on those efforts soon. But the need to work harder is clear for all to see.

Reward synthesis

Enlarge and incentivize efforts that examine past discoveries.

he idea that scientists can see to make discoveries only because they stand on the shoulders of giants was popularized by Isaac Newton. Fittingly, he borrowed the idea from a significant figure who had gone before him, probably the twelfth-century French philosopher Bernard of Chartres. It's a sound principle: build on previous efforts to seek and find truth. But a vast number of previous discoveries are now captured in an overwhelmingly large body of literature — so what is a modern truth-seeker to do?

One strategy is to distil knowledge in a way that empowers those needing to resolve a practical solution. This comes in many forms — from the regular, heroic efforts of the Intergovernmental Panel on Climate Change to the ad-hoc assessments done urgently to help steer decisions on political or environmental crises. Other efforts — largely by not-for-profit organizations committed to evidence-based assessment — are driven by a need to ensure that the best possible outcomes will follow any intervention. The Cochrane reviews (go.nature.com/2jqocex) exemplify this assessment for best practice in health contexts, and the Campbell reviews (go.nature.com/2k86p1p)

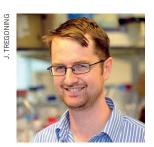
do so in social, educational and behavioural contexts.

As societal challenges grow in research priority, there is ever more need for such synthesis. But it takes effort, as described by, for example, a Cornell University Library guide to a systematic review (go.nature. com/2k6ftil). And, more problematically, the academic ecosystem does not incentivize such work.

To help nudge the system in that direction, *Nature* this week publishes two Comment articles that highlight the importance of such assessments of evidence, and suggest ways to maximize their effectiveness. In the first (page 361), several experts from policy, funding and publishing (including *Nature*'s editor-in-chief) present four principles to help make evidence syntheses aimed at policymakers easier to commission, and more powerful in delivery and implementation. The second (page 364), by two researchers who focus on evidence for conservation biology, discusses a form of evidence synthesis that can provide a more cost-effective way to appraise evidence when data are sparse and patchy. This is a reflection of the reality that, for some interventions, randomized controlled trials aren't possible, but there is, nevertheless, a need to make sense of the available evidence.

More scientists should identify fields for which such an exercise is necessary (or will be soon) and, after proper consultation with policymakers on what questions are most relevant, they should produce a useful assessment of the evidence. We hope that these articles will encourage researchers, and their institutions, funders and publishers, to recognize the benefits that good syntheses of knowledge will provide. \blacksquare

WORLD VIEW A personal take on events



How will you judge me if not by impact factor?

Stop saying that publication metrics don't matter, and tell early-career researchers what does, says **John Tregoning**.

R umours among junior faculty members are that reports of the death of the impact factor are greatly exaggerated. In a round of funding earlier this year, my research output was described as being in "high-impact journals" by one reviewer and in "middle-tier journals" by another, with knock-on effects on their grant scores. It is not unheard of for people to be told that the only articles that count are the ones in journals with an impact factor that is over an arbitrary value. Or, worse, that publishing in low-tier journals pollutes their CVs.

That's true even at institutions that have signed on to the San Francisco Declaration on Research Assessment (DORA), which advocates replacing journal impact factors (JIFs) with something better and fairer.

Actually, pretty much everyone agrees that the use of the journal impact factor as the sole tool to evaluate research is a bad thing.

But for all the invective heaped on the JIF as a metric, no alternative has emerged. The activation energy to find something else is just too high. The JIF is wrong in so many ways, but it is so easy, a number that lets you rank scientists and their output in the same way as experimental data. It is also quick — scanning a list of journals takes very little time — and deeply ingrained. Also, when viewed macroscopically, it's not entirely wrong. Papers published in journals with higher impact factors tend, on average, to be better and more important than those in journals with lower ones.

We are told that the impact factor should no longer be used, but not told what to use instead. So where does that leave the early-career

researcher eyeing the conventional academic track? Straddling uncertainty and the status quo. And stressed out and less productive as a result.

Ideally, just putting our research out there should be enough for people to descry our brilliance and promote it accordingly. But that is not how the system works.

My peers who have focused on getting articles in high-impact journals seem to have outperformed those with better social-media presence. But I am judging their success in part by their ability to publish papers in high-impact journals!

It's no secret that doing great science does not necessarily overlap with having a great career. The current system masquerades as a meritocracy, but it is subjective, biased, built on personal networks and laced with blind luck. To succeed, we need to leverage our reputation, and the main tool we have for this is our research output. So we need to be strategic about where we place our work, to ensure that the right people notice it. To stay competitive, we need a map and time to navigate it.

The impact factor used to provide that map. For people with few publications, the nice thing about JIFs is that they are prospective rather than retrospective. JIFs give an instant validation; both h-index and citations

increase over time, a luxury that early-career researchers have not yet accumulated. With the old system, if you worked hard, you got your first-author papers in journals with a high impact factor. That brought tenure, keys to the executive toilet and (an ancient principal investigator once promised) lifelong happiness. Sure, this was a fickle route that favoured trainees who were lucky enough to find the rare laboratory with an on-ramp to the fast track. But at least we spent less time feeling lost.

Now, who knows what counts? What about that abstract that I naively submitted to a predatory journal when trying to get someone to pay for a trip to the United States? How does that tot up against a full paper in this journal? Or this spunky essay?

Although DORA is in my heart, impact factors are still on my mind.

Of course, I have some broad-brush suggestions to throw into the mix. There should be more than one route to your destination. There should also be more than one destination. We need to find ways to rate and recognize our broader contribution to the community (including public engagement, internal committees and teaching). In fact, the rising generation of scientists has a unique set of strengths that could make for a stronger scientific enterprise in the long term, if hiring committees thought to reward it. Those attributes include a more socially networked approach to doing science, plus a facility to use information technology to share data, methods and credit. Hiring and reviewing committees need to recognize that the old system is not ideal for selecting future

scientific leadership.

Right now, however, I don't think there is a need for more ideas about how to overhaul the game. I want more clarity on what the game actually is. It doesn't have to be universal; it does have to be transparent. If different institutions play by different rules, that's okay — I'll work out a way to play to my strengths. But it is difficult to play when you don't know the rules, harder still when the rules change each time you look for a new position. At this point, I would settle for impact factor as the least-bad option; at least it's something.

Maybe the DORA advocates will figure out a fantastically fair way to gauge scientific output in five years, or ten. Maybe it will be holistic, broadly accepted, supportive and simple. That would be great, when it happens.

In the meantime, confusion over how to judge scientific productivity is sapping scientific productivity. We need a quick fix, and the quickest fix is clarity. \blacksquare

John Tregoning is a senior lecturer at Imperial College London, where he studies the immune response to viral infections. e-mail: john.tregoning@imperial.ac.uk



SEVEN DAYS The news in brief

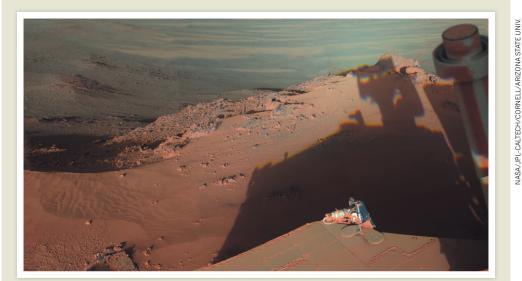
EVENTS

Water shortage

India is experiencing the worst water crisis in its history. according to a government study. The report, released on 14 June, examined water supplies in 24 of the country's 36 states and union territories and found that nearly 70% of fresh water is contaminated. The report also found that groundwater resources — which account for 40% of the country's freshwater supply — are being depleted at unsustainable rates. By 2020, 21 major cities are expected to run out of groundwater. Poor rainwaterstorage infrastructure is contributing to the crisis. The government agency in charge of water in New Delhi says the city loses 40% of its drinking water from leaking pipes and unauthorized connections. A 2016 report by non-profit group WaterAid found that an estimated 76 million people in India have no access to safe drinking water.

Study cancelled

On 15 June, the US National Institutes of Health (NIH) terminated a controversial US\$100-million study into whether drinking small amounts of alcohol every day can improve health. The agency's decision came shortly after an NIH advisory council voted unanimously to end the trial, dubbed MACH15. An NIH investigation had found that agency staff and outside researchers had acted inappropriately by soliciting industry funding and biasing the grant-review process. "We are deeply disappointed that issues raised have led to a recommendation to end the trial," said lead investigator and cardiovascular researcher Kenneth Mukamal of



Mars dust endangers NASA rover

A huge dust storm is blanketing much of Mars, blocking the sunlight that NASA's 15-year-old Opportunity rover needs to survive. Mission controllers have not heard from the solarpowered rover since 10 June, and think it is in a low-power mode in which everything except its clock is turned off. If the rover's power level

and temperature don't drop too low - and predictions suggest they won't — then it might be able to wake itself once the dust has cleared. That could take weeks. Opportunity landed on Mars in January 2004, and was designed to last 90 Martian days, or about 13 Earth weeks. See go.nature.com/2mdihk2 for more.

Beth Israel Deaconess Medical Center in Boston. Massachusetts, in a statement. "We strongly believe that MACH15 has a critically important scientific premise, rigorous design, and highlyqualified team," he says. See go.nature.com/2ma8lrh for more.

Harassment study

Sexual harassment is pervasive throughout academic science in the United States, driving talented researchers out of the field and harming others' careers, finds a report from the US National Academies of Sciences, Engineering, and Medicine in Washington DC. The analysis concludes that policies to fight the problem are ineffective because they are set up to protect institutions, not victims — and that universities, funding agencies,

scientific societies and other organizations must take stronger action. The report, released on 12 June, is the most comprehensive study yet on the extent of harassment in the sciences. See page 352 for more

Diet debacle

The New England Journal of Medicine retracted a landmark nutritional study of the Mediterranean diet and published a corrected article on 13 June. The study, which evaluated the diets of 7,447 participants for a median of 4.8 years, is one of six that the journal has corrected in the wake of a 2017 analysis that flagged possible statistical abnormalities (J. B. Carlisle Anaesthesia 72, 944-952; 2017). Despite the revisions, the corrected version of the Mediterranean

diet study still concluded that supplementing the diet with extra-virgin olive oil or nuts can make it beneficial for those at high risk of heart disease (R. Estruch et al. N. Engl. J. Med. http://doi.org/ cq2s; 2018).

TECHNOLOGY

DeepMind ethics

Health-technology firm DeepMind Health has agreed to abide by 12 ethics principles proposed by an independent panel that regularly reviews its work. The London company is owned by Google's parent firm Alphabet, and is developing tools that use artificial intelligence (AI) to improve health services. The principles, published on 15 June, include making sure that the patients whose data are used to create the company's algorithms

348 | NATURE | VOL 558 | 21 JUNE 2018

benefit from them, and that the firm does not become a monopoly or make "excessive profits". The panel's annual review report adds that DeepMind Health should specify and publicly disclose how it proposes to make money in the future. This month, Google also unveiled broader rules governing its AI usage.

Al ethics code

The Singapore government will develop a voluntary code of ethics for businesses that use artificial intelligence (AI) and personal data. The code will focus on ensuring that AI use in industry is transparent and fair. The country will appoint an advisory council, with members from both the public and private sectors, to oversee the plan. A five-year research programme at the Singapore Management University will also examine the ethical, legal, policy and governance issues arising from AI and data use. The city-state joins countries such as India and the United Kingdom in exploring ethical AI use.

PEOPLE

Science adviser

The New Zealand government appointed biochemist Juliet Gerrard as its next chief scientific adviser on 12 June. Gerrard (pictured) will be the second person and first



woman to occupy the post, which advises the prime minister on science and science-policy issues. She is a faculty member at the University of Auckland whose research focuses on how the structure and assembly of a protein influence its function. Gerrard said she hopes her appointment will encourage young female scientists to be ambitious in their careers. She will take over from Peter Gluckman — who was appointed in 2008 — on 1 July. See go.nature.com/2t95fcl for

Fraud charges

On 14 June, Elizabeth Holmes and Ramesh Balwani, respectively the former chief executive and former president of health-technology company Theranos in Palo Alto, California, were each charged with conspiracy to commit wire fraud and with wire fraud. According to the indictment from the US attorney's office in San Jose, California, Holmes

and Balwani used electronic communications to mislead physicians, patients and investors with false claims about the accuracy and reliability of the company's fast and cheap blood analyser, and about Theranos's financial condition. Balwani's lawyer, Jeffrey Coopersmith, said in a statement that his client is "innocent, and looks forward to clearing his name at trial". Theranos declined to comment on the case but announced that Holmes was no longer chief executive, although she remains chair of the board. Holmes did not respond to a request for comment.

POLICY

Delayed rule

The US Department of Health and Human Services (HHS) has delayed implementing reforms to the 'Common Rule, which governs research involving human subjects. On 18 June, the HHS said that institutions had until 21 January 2019 to comply with the latest version of the Common Rule, which aims to streamline regulatory safeguards such as ethics approval boards and consent forms for storing biological samples. It is the second delay for the revised regulation, which was initially scheduled to take effect in January 2018. Universities and medicalresearch institutions had

lobbied for more time to institute the changes and clarify the rule's requirements.

FACILITIES

Particle collider

Work on a major upgrade to the world's most powerful particle collider began on 15 June. The Large Hadron Collider at CERN, Europe's particle-physics lab near Geneva, Switzerland, began smashing together protons in its 27-kilometre-long accelerator ring a decade ago, to test fundamental physics theories and hunt for new particles. It can currently collide about 1 billion protons per second. Once the work is complete in 2026, the upgraded 'high luminosity' machine will use more-powerful magnets to generate up to 7 times as many collisions and collect 10 times as much data. Finer data will allow physicists to make moreaccurate measurements and to see the influence of as-yetundiscovered phenomena.

PUBLISHING

US open-access deal

A US university has signed an agreement with an academic publisher to offset the costs of open-access publishing against journal subscriptions. The deal between the Massachusetts Institute of Technology in Cambridge and the Royal Society of Chemistry, a British scholarly publisher, is the first of its type in North America, according to a statement released by the university on 14 June. Some proponents of the contracts, known as read-and-publish agreements, say they are an important step towards removing paywalls on some publicly funded research articles. Several consortia that negotiate publishing agreements on behalf of universities in European countries have agreed read-and-publish deals in recent years (see go.nature. com/2thwb4r).

◇ NATURE.COM

For daily news updates see:

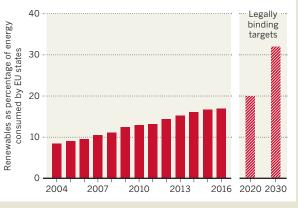
www.nature.com/news

TREND WATCH

European Union policymakers reached a deal on 14 June to strengthen the bloc's legally binding targets for renewable energy. By 2030, 32% of energy consumed in the EU should be from renewable sources such as wind and solar, up from the previous goal of 27%. An interim target, to obtain 20% of energy from clean sources by 2020, remains in place. The deal includes phasing out the use of palm oil — widely blamed for contributing to deforestation — as a transport fuel by 2030.

EUROPE'S RENEWABLE-ENERGY GOALS

Under a new legally binding target, the European Union will need to obtain 32% of its energy from renewables by 2030; in 2016, renewables accounted for 17% of energy consumed.



NEWSINFOCUS

LAB LIFE Sexual harassment is common in US academic science, report finds p.352

BUSINESS Microsoft's purchase of Github worries researchers **p.353**

MEDICAL RESEARCH The hunt is on for dormant cancer cells p.355





Security measures implemented in advance of the World Cup have affected some scientists in Russia.

PULICY

World Cup chemical ban frustrates Russian labs

Sports event adds to systemic resupply problems for biochemists in the country.

BY QUIRIN SCHIERMEIER

Russia on 14 June. But some Russian researchers might find themselves with more time to watch the matches than they expected.

Because of security and counter-terror measures enacted by the government ahead of the World Cup tournament, some Russian labs will go without the radioactive reagents that they urgently need for their research, according to molecular biologists and biochemists who spoke to *Nature*.

In a presidential decree issued on 11 May, the Russian government suspended the sale and transport of hazardous chemical and biological substances — including toxic and radioactive chemicals — for two months, citing security concerns. The World Cup runs until 15 July. The decree applies only to cities hosting the matches, but many of these, including Moscow, happen to be research hubs, says Konstantin Severinov, a biochemist at the Skolkovo Institute of Science and Technology (Skoltech) near Moscow.

The measures threaten to stall the relatively little molecular-biology research that exists in

Russia, says Severinov. Last month, Russian researchers who had recently ordered radioactive nucleotides, which they use to measure gene expression and for other assays, received bad news from the Russian Academy of Sciences' Institute of Bioorganic Chemistry in Moscow: an expected June delivery to their labs would be cancelled because of the presidential decree. No other Russian centre supplies such reagents.

"This jeopardizes the whole workflow in my lab," says Severinov, who is also a group leader at the Russian Academy of Sciences' Institutes of Molecular Genetics and Gene Biology in

▶ Moscow. Numerous projects — including CRISPR—Cas9 gene-editing experiments and those measuring the effects of toxins on cells — have been affected, he says.

Maintaining supplies of research reagents and other consumables is notoriously problematic in Russia, says Stephen O'Brien, director of the Theodosius Dobzhansky Center for Genome Bioinformatics in Saint Petersburg. Russian production capacities are slight, and severe customs restrictions effectively bar scientists who depend on radio-labelled reagents from legally purchasing them from foreign suppliers, Severinov says.

DOMESTIC DEMAND

Meanwhile, domestic supply is routinely hampered by bureaucracy and long delivery times. "We always have problems with ordering research materials during summer," says Ilya Osterman, a biochemist at the Skoltech Center for Translational Biomedicine in Moscow, who uses restricted chemicals to examine the shapes of different RNA molecules and to measure gene expression. "The World Cup only makes the situation worse."

To prevent frustrating disruptions to their research, scientists in Russia must order such reagents several weeks in advance, through their institution's procurement department. With the World Cup and the ensuing summer break, the next deliveries of radio-labelled nucleotides might not arrive until early autumn. "This means a bad disruption," says Severinov. "Four of my PhD students are caught midway in their thesis work."

Alexei Khokhlov, a vice-president of the Russian Academy of Sciences, which runs the institute that supplies researchers with radio-labelled nucleotides, did not reply to an e-mail from *Nature* asking how many scientists were affected and how the delay might affect their research.

Before his re-election as president in March, Vladimir Putin promised to strengthen Russia's struggling research base. But strict customs and import restrictions on research materials continue to put Russian scientists at a competitive disadvantage compared with researchers in countries where there is an ample supply of chemicals and science equipment, says Fyodor Kondrashov, a Russian biologist at the Institute of Science and Technology Austria in Klosterneuburg.

The enhanced security restrictions will be lifted soon after the World Cup final takes place at the Luzhniki Stadium in Moscow on 15 July. "This current crisis might be short-lived," says Kondrashov. "But it underlines the difficulty of doing cutting-edge research in a country that is not entirely free."

LAB LIFE

Sexual harassment is rife in US science

Science academies call for cultural shift to fight problem.

BY ALEXANDRA WITZE

exual harassment is pervasive throughout academic science in the United States, driving some talented researchers out of the field and harming others' careers, finds a report from the US National Academies of Sciences, Engineering, and Medicine in Washington DC. The analysis concludes that policies to fight the problem are ineffective because they are set up to protect institutions, not victims — and that universities, funding agencies, scientific societies and other organizations must take stronger action.

"The cumulative effect of sexual harassment is extremely damaging," says Paula Johnson, president of Wellesley College in Massachusetts and co-chair of the committee that wrote the report. "It's critical to move beyond the notion of legal compliance to really addressing culture."

The report, released on 12 June, is the most comprehensive look yet at harassment in the sciences. It comes in the wake of the #MeToo movement against sexual assault and harassment, and as the US national academies are grappling with whether to punish members accused of

harassment.

Notably, the report finds that the main mechanism for reporting sexual

"It's not okay to treat your co-workers like dirt."

harassment on US campuses — Title IX, the federal law enacted in 1972 that outlaws discrimination on the basis of gender — has not reduced the incidence of sexual harassment. Institutions can find ways to comply with Title IX that avoid liability but don't actually prevent harassment, says Asmeret Asefaw Berhe, a biogeochemist at the University of California, Merced.

To change this, the report says, research institutions should act to reduce the power differential between students and faculty members, perhaps by introducing group-based advising; the government should prohibit confidentiality in settlement agreements, so that harassers cannot switch jobs without their new employer knowing about past behaviour; and research organizations should treat sexual harassment at least as seriously as research misconduct.

"This is an incredibly comprehensive and ambitious report," says Anna Bull, a sociologist

at the University of Portsmouth, UK, and co-founder of The 1752 Group, which works to end harassment in academia. "They get beyond the 'one bad apple' approach and look at the culture that enables that one bad apple."

The most common type of sexual harassment is gender harassment, the report says. Such behaviour conveys the idea that women don't belong in the workplace or merit respect — "the put-downs as opposed to the come-ons," Johnson says. Such actions might seem minor but can seriously affect the person targeted, she adds; they also set the stage for unwanted sexual attention and coercion.

TRACKING THE TOLL

All three kinds of sexual harassment are illegal in the United States when they interfere with a person's work environment, yet all are widespread in science, engineering and medicine. Previous research has shown that the prevalence of reported sexual harassment in US academia, at 58%, is second only to the military's 69%, and outpaces that of industry and government¹. Women of colour experience particularly high rates of harassment², as do people from sexual- and gender-minority groups³.4. Men in academia also experience sexual harassment, although at lower rates than women do³.

To build on those earlier studies, the academies' committee commissioned an analysis that found that 20% of female science students at the University of Texas's campuses reported being sexually harassed by faculty members or staff there. A similar survey of the Pennsylvania State University system concluded that 43% of graduate students experienced harassment (see 'Pervasive problem').

All types of harassment, including gender harassment, can prove corrosive to scientists' career development, according to interviews of 40 women faculty members conducted for the new report. One woman who had been raped by a colleague gave up research; another, who had been verbally berated by her dean, felt the experience derailed her from ever becoming a full professor.

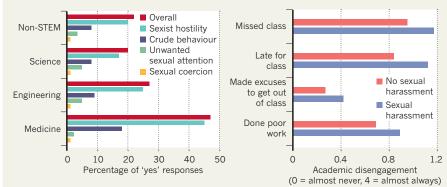
"It's not okay to treat your co-workers like dirt," says Kathryn Clancy, a biological anthropologist at the University of Illinois at Urbana–Champaign and a member of the report committee. But university leaders often minimized or ignored the harassing behaviour, survey participants said, especially when it involved higher-ranking faculty members who

PERVASIVE PROBLEM

All forms of sexual harassment are prevalent in US academic science, a new report finds.

Harassment by major. The proportion of female students in the University of Texas system who report having been harassed by faculty members or staff varies between those who major in science, technology, engineering and medicine (STEM) and those who do not.

Academic impact. Female science majors at the University of Texas who say they have been harassed by faculty members or staff also report higher rates of disengagement with their studies.



were perceived as stars in their department.

An institution's workplace climate is by far the greatest predictor of sexual harassment, the academies' report says. Title IX and related laws are a good start, says Clancy, but universities need to embrace other methods of addressing sexual harassment. These include ways for victims to report incidents without being re-traumatized or subjected to retaliation.

"Many targets of harassment are women and minorities in vulnerable positions," says Akiko

Iwasaki, an immunologist at Yale University in New Haven, Connecticut. "If they feel like their careers rely on future recommendation letters from the harassers, they are less likely to want to come forward."

However strong the report's findings, it is still up to universities to interpret them, says Jessica Cantlon, a cognitive neuroscientist who is in the process of leaving the University of Rochester in New York. There, she was part of a group of faculty members who sued the university over its handling of sexualharassment allegations against a researcher in her department; the case is ongoing. "We are still waiting for tangible changes at our university, despite having voiced similar recommendations over two years ago in the wake of multiple student complaints about sexual harassment by a faculty member," she says.

The report comes as the flagship national academy is facing criticism over its policies on harassment. Since early May, more than 3,500 people have signed a petition requesting that the National Academy of Sciences expel members who have been sanctioned for sexual harassment, retaliation or assault.

Academy president Marcia McNutt says the group's governing council will consider proposed changes when it meets in August. "This is something we have to take seriously as an organization," she says. But, she adds, the academy would probably not initiate its own investigation of a member — instead referring any complaints that it receives to the leadership of that person's university. "One is ongoing right now," she says. "No, I won't tell you who it is." ■

- Ilies, R., Hauserman, N., Schwochau, S. & Stibal, J.
- Personnel Psychol. **56**, 607–631 (2003). Clancy, K. B. H., Lee, K. M. N., Rodgers, E. M. & Richey, C. J. Geophys. Res. Planets 122, 1610-1623
- Konik, J. & Cortina, L. M. Social Justice Res. 21, 313-337 (2008).
- Rosenthal, M. N., Smidt, A. M. & Freyd, J. J. Psychol. Women Q. 40, 364-377 (2016).

SOURCE: NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE

Microsoft's GitHub buyout raises fears

Users worry popular data-sharing site will become less open.

BY ANDREW SILVER

■ itHub — a website that has become popular with scientists collaborating on research data and software — is to be acquired by Microsoft for US\$7.5 billion. In the wake of the takeover announcement on 4 June, some scientists and programmers voiced concerns about the deal on social media. They fear that the site will become less open, or less useful for sharing and tracking scientific data, after the buyout. But others are hopeful that Microsoft's stewardship will make the platform even more valuable.

GitHub launched in 2008, and is now widely used to store, share and update data sets and software code. As of 13 June, more than 223,000 academic papers on Google Scholar cited the website, which is free to use for projects that release their code. GitHub uses a version-control software known as Git, which transparently records changes to files. This allows programmers in different locations to work on the same project in real time, and to track changes and merge updated data.

Although Microsoft says GitHub will remain open to any project, some scientists are sceptical about that commitment. "Open Science is not compatible with one corporation owning the platform used to collaborate on code. I hope that expert coders in #openscience have a viable alternative to #github," tweeted Tom Johnstone, a cognitive neuroscientist at the University of Reading, UK.

Björn Grüning, a bioinformatician at the University of Freiburg in Germany, says some

researchers are wary because Microsoft has been slow to make its own tools available in open-source code, and to make its services compatible with open-source projects. He has several projects on GitHub, but says he will move them to another service if the company makes the platform less open, forces Microsoft tools on users or changes its pricing model.

Mahmood Zargar, who studies opensource communities at the Free University of Amsterdam, is more concerned that Microsoft will impose changes that will make GitHub less efficient for him to use. He's planning to move his projects to other services.

A spokesperson for Microsoft did not answer Nature's questions about researchers' concerns, but referred to a blogpost by company chief executive Satya Nadella. "We are committed to being stewards of the GitHub community, which will retain its developer-first ethos, operate independently and remain an open platform," Nadella wrote.

Arfon Smith, a data-science manager at the Space Telescope Science Institute in Baltimore, Maryland, says the fears are overblown. He doesn't think Microsoft will change the features that researchers care about, such as its ease of use. Katy Huff, a nuclear engineer at the University of Illinois at Urbana-Champaign, thinks GitHub will give Microsoft an opportunity to support science.

Expanded human gene tally reignites debate

After 15 years, researchers still can't agree on how many genes are in the human genome.

BY CASSANDRA WILLYARD

ne of the earliest attempts to estimate the number of genes in the human genome involved tipsy geneticists, a bar in Cold Spring Harbor, New York, and pure guesswork.

That was in 2000, when a draft human genome sequence was still in the works; geneticists were running a sweepstake on how many genes humans have, and wagers ranged from tens of thousands to hundreds of thousands. Almost two decades later, scientists armed with real data still can't agree on the number a knowledge gap that they say hampers efforts to spot disease-related mutations.

The latest attempt to plug that gap uses data from hundreds of human tissue samples and was posted on the bioRxiv preprint server on 29 May (M. Pertea et al. Preprint at bioRxiv http://doi.org/cq5s; 2018). It includes almost 5,000 genes that haven't previously been spotted — among them nearly 1,200 that carry instructions for making proteins. And the overall tally of more than 21,000 proteincoding genes is a substantial jump from previous estimates, which put the figure at around 20,000.

GENE TALLY

But many geneticists aren't yet convinced that all the newly proposed genes will stand up to close scrutiny. Their criticisms underscore just how difficult it is to identify new genes, or

"People have been working hard at this for 20 years, and we still don't have

even to define what a gene is.

"People have been working hard at this for 20 years, and we still don't have the answer," says Steven Salzberg, a computa-

tional biologist at Johns Hopkins University in Baltimore, Maryland, whose team produced the latest count.

HARD TO PIN DOWN

the answer."

In 2000, with the genomics community abuzz over the question of how many human genes would be found, researcher Ewan Birney launched the GeneSweep contest. Birney, now co-director of the European Bioinformatics Institute (EBI) in Hinxton, UK, took the first bets at a bar during an annual genetics meeting, and the contest eventually attracted more than 1,000 entries and a US\$3,000 jackpot. Bets on the number of genes ranged from more than 312,000 to just under 26,000, with an average of around 40,000. These days, the span of estimates has shrunk — with most now between 19,000 and 22,000 — but there is still disagreement (see 'Gene tally').

Salzberg's team used data from the Genotype-Tissue Expression (GTEx) project, which sequenced RNA from more than 30 different tissues taken from several hundred cadavers. RNA is the intermediary between DNA and proteins. The researchers wanted to identify genes that encode a protein and those that don't, but that still have an important role in cells. So they assembled GTEx's 900 billion tiny RNA snippets and aligned them with the human genome.

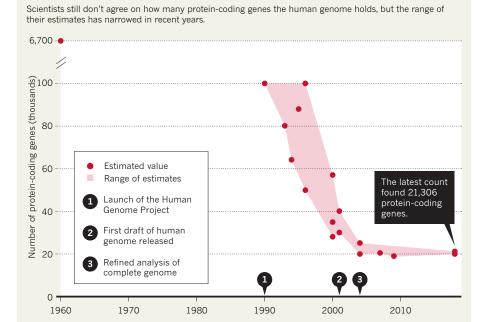
Just because a stretch of DNA is expressed as RNA, however, does not necessarily mean it's a gene. So the team attempted to filter out noise using a variety of criteria. For example, the researchers compared their results with genomes from other species, reasoning that sequences shared by distantly related creatures have probably been preserved by evolution because they serve a useful purpose, and so are likely to be genes.

The team was left with 21,306 proteincoding genes and 21,856 non-coding genes — many more than are included in the two most widely used human-gene databases. The GENCODE gene set, maintained by the EBI, g includes 19,901 protein-coding genes and 15,779 non-coding genes. RefSeq, a database 🕏 run by the US National Center for Biotechnol- ຜ ogy Information (NCBI), lists 20,203 protein- $\stackrel{\circ}{\leq}$ coding genes and 17,871 non-coding genes.

Kim Pruitt, a genome researcher at the NCBI in Bethesda, Maryland, and a former 💆 head of RefSeq, says the difference is probably due in part to the volume of data that Salzberg's team analysed. RefSeq relies on an older data set that contains 21 billion short sequences. GENCODE uses different data again: a type that makes recognizing transcripts easier, but which can miss genes. And there's another major difference. Both GENCODE and Ref-Seq use manual curation — a person reviews the evidence for the gene and makes a final determination. Salzberg's group relied solely on computer programs to sift the data.

"If people like our gene list, then maybe a couple years from now we'll be the arbiter of human genes," says Salzberg.

But many scientists say they need more evidence to be convinced that the latest list is



accurate. Adam Frankish, a computational biologist at the EBI who coordinates the manual annotation of GENCODE, says that he and his group have scanned about 100 of the protein-coding genes identified by Salzberg's team. By their assessment, only one of those seems to be a true protein-coding gene. And Pruitt's team looked at about a dozen of the Salzberg group's new protein-coding genes, but didn't find any that would meet RefSeq's criteria.

Salzberg acknowledges that the new genes on his team's list will require validation by his group and others.

Further confounding counting efforts is the imprecise and changing definition of a gene. Biologists used to see genes as sequences that code for proteins, but then it became clear that some non-coding RNA molecules have important roles in cells. Judging which are important — and should be deemed genes — is controversial, and could explain some of the discrepancies between Salzberg's count and others.

Having an accurate tally of all human genes is key for efforts to uncover links between

genes and disease. Uncounted genes are often ignored, even if they contain a disease-causing mutation, Salzberg says. But hastily adding genes to the master list can pose risks, too, says Frankish. A gene that turns out to be incorrect can divert geneticists' attention away from the real problem.

Still, the inconsistencies in the number of genes from database to database are problematic for researchers, Pruitt says. "People want one answer," she adds, "but biology is complex."

MEDICAL RESEARCH

Silent cancer cells targeted

Researchers hunt dormant cells that break off tumours, and aim to keep them asleep.

BY HEIDI LEDFORD

A fter decades of designing drugs to kill rapidly dividing tumour cells, many cancer researchers are switching gears: targeting malignant cells that lie silent and scattered around the body, before they give rise to new tumours.

These cells seed the metastases responsible for about 90% of cancer deaths. They are the source of the heartbreaking cancer resurgence seen in many people whose seemingly successful initial treatment had fostered hopes that they were cured. Treatments that target proliferating tumour cells often miss these silent cells because they're not actively dividing.

Dormant cancer cells are rare, and they are difficult to sift from the trillions of normal cells in the body. For years, scientists lacked the tools to study them, says cancer researcher Julio Aguirre-Ghiso of the Icahn School of Medicine at Mount Sinai in New York City. But that is beginning to change.

From 19 to 22 June, researchers will gather in Montreal, Canada, for what Aguirre-Ghiso says is the first meeting dedicated to these sleeper cancer cells. "The mass of investigators has reached a critical number," he says. "And there is the realization that it's an important clinical need."

That demand is particularly acute in cancers

— such as those in the breast, prostate and pancreas — that recur at a high rate, sometimes many years after treatment. "You remove the tumour, you irradiate, you do this, you do that," says cancer researcher Mina Bissell, of the Lawrence Berkeley National Laboratory in California. "But sooner or later the cancer metastasizes, and you say to yourself, 'Where did these things come from?"

CELL SPOTTING

Mounting evidence suggests that dormant cells break away from a parent tumour early in its development and travel through blood vessels to new sites in the body (see *Nature Methods* 15, 249–252; 2018). But then, after settling into other tissues or organs, such cells will effectively go to sleep, lying dormant until a trigger — as yet unknown — rouses them. Only then do they begin dividing and form a new tumour.

When cancer researchers tried to study this dormancy, they quickly ran into a problem: mouse models of cancer had been designed to generate quick-growing and highly lethal parent, or primary, tumours. Researchers studying dormancy, however, need slow-growing tumours — which have time to shed rogue cancer cells — and the ability to track those cells long after the primary tumour has been removed.

"Those sorts of animals have been very difficult to develop," says Kathy Miller, a

breast-cancer specialist at Indiana University in Indianapolis. But several labs have made progress, developing models to track dormant cells in mice for more than a year.

Techniques for identifying those cells are also improving. Joshua Snyder, a cell biologist at Duke University School of Medicine in Durham, North Carolina, uses a mix of fluorescent markers to identify and trace rogue cells expressing cancer-linked genes.

"As long as those cells remain dormant, they're not killing my patient."

And at the meeting in Montreal, geneticist Jason Bielas of the Fred Hutchinson Cancer Research Center in Seattle, Washington, will present

preliminary results from his efforts to barcode such cells using specific DNA sequences. The cells can then be identified using cheap DNA-detection methods at a resolution of about one in one billion cells.

IDENTIFYING INHIBITIONS

Once the silent cells are identified, new methods for determining which genes they express could help researchers to pin down the factors that induce dormancy and the triggers that can rouse sleeping cells. With that information, it might be possible to prevent the cells from waking, says Miller. "As long as those cells remain





Save time get the Nature Briefing direct to your inbox every day go.nature.com/ savetime

MORE NEWS

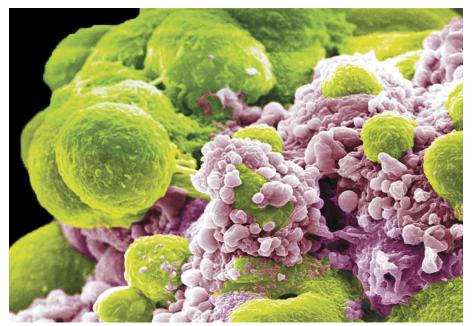
- Controversial US alcohol study cancelled go.nature.com/2mca0gg
- Tech entrepreneur doubles down on critique of NASA mission go.nature.com/2yn2vgh
- Mammals turn to night life to avoid people go.nature.com/2lj7e35

NATURE PODCAST



Pancreatic cancer weight loss; tiny silicon cages; and bias in Al algorithms

nature.com/nature/



Cancer cells (green) can splinter off of tumours and settle in other parts of the body.

▶ dormant, they're not killing my patient."
Efforts to keep sleeper cells at bay are also under way. In 2015, Aguirre-Ghiso's laboratory reported that a combination of two approved drugs — 5-azadeoxycytidine and retinoic acid — could induce dormancy in prostate-

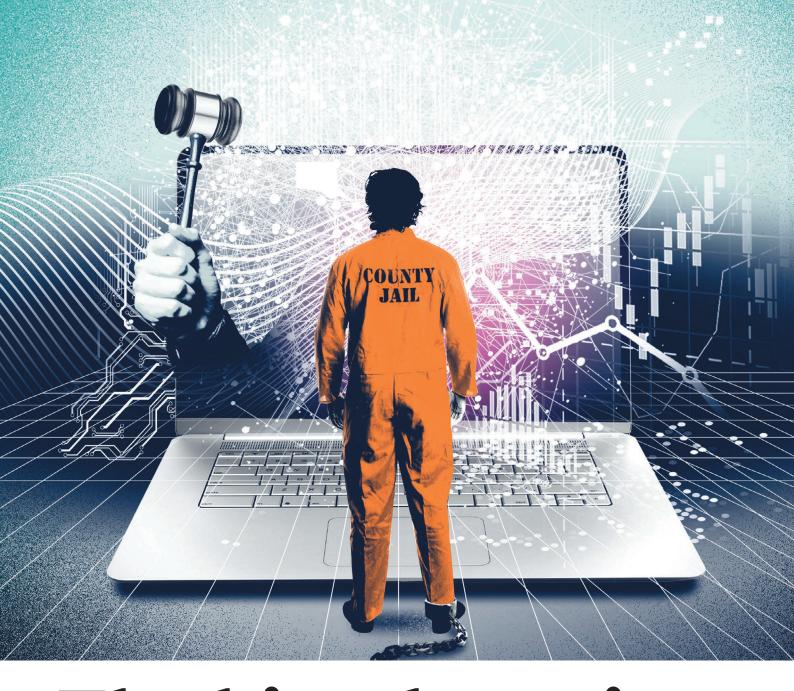
cancer cells grown in culture, as well as in mice (M. S. Sosa *et al. Nature Commun.* **6**, 6170; 2015). Now, William Oh, an oncologist at Mount Sinai, and his colleagues are enrolling people with prostate cancer in a trial to test these findings in the clinic.

Others are looking at ways to kill the dormant cells outright. Cancer researcher Veronica Calvo-Vidal and her colleagues in Aguirre-Ghiso's lab have collaborated with pharmaceutical firm Eli Lilly of Indianapolis to characterize an inhibitor of a protein called PERK, which is expressed at unusually high levels in dormant cancer cells. Early studies in mice suggest that the inhibitor can kill the cells, and the team is now analysing gene expression in individual dormant cells to learn more about how the molecule works.

But Miller cautions that it is also important to develop ways of identifying which cancers are most likely to recur, so that physicians can select the patients who warrant such treatments. She and other oncologists who treat breast cancer already struggle to decide which patients should receive further hormone therapy to reduce the risk of recurring tumours. "We are moving closer to the day when we are able to do a much better job of identifying recurrence much earlier," she says.

CORRECTION

The News Feature 'Here come the waves' (*Nature* **556**, 164–168; 2018) incorrectly described llya Mandel as a LIGO theorist. Mandel left the LIGO collaboration in 2016.



The bias detectives

As machine learning infiltrates society, scientists grapple with how to make algorithms fair.

BY RACHEL COURTLAND

In 2015, a worried father asked Rhema Vaithianathan a question that still weighs on her mind. A small crowd had gathered in a basement room in Pittsburgh, Pennsylvania, to hear her explain how software might tackle child abuse. Each day, the area's hotline receives dozens of calls from people who suspect that a child is in danger; some of these are then flagged by call-centre staff for investigation. But the system does not catch all cases of abuse. Vaithianathan and her colleagues had just won a half-million-dollar contract to build an algorithm to help.

Vaithianathan, a health economist who co-directs the Centre for Social Data Analytics at the Auckland University of Technology in New Zealand, told the crowd how the algorithm might work. For example, a tool trained on reams of data — including family backgrounds and criminal records — could generate risk scores when calls come in. That

could help call screeners to flag which families to investigate.

After Vaithianathan invited questions from her audience, the father stood up to speak. He had struggled with drug addiction, he said, and social workers had removed a child from his home in the past. But he had been clean for some time. With a computer assessing his records, would the effort he'd made to turn his life around count for nothing? In other words: would algorithms judge him unfairly?

Vaithianathan assured him that a human would always be in the loop, so his efforts would not be overlooked. But now that the automated tool has been deployed, she still thinks about his question. Computer calculations are increasingly being used to steer potentially life-changing decisions, including which people to detain after they have been charged with a crime; which families to investigate for potential child abuse,

and — in a trend called 'predictive policing' — which neighbourhoods police should focus on. These tools promise to make decisions more consistent, accurate and rigorous. But oversight is limited: no one knows how many are in use. And their potential for unfairness is raising alarm. In 2016, for instance, US journalists argued that a system used to assess the risk of future criminal activity discriminates against black defendants.

"What concerns me most is the idea that we're coming up with systems that are supposed to ameliorate problems [but] that might end up exacerbating them," says Kate Crawford, co-founder of the AI Now Institute, a research centre at New York University that studies the social implications of artificial intelligence.

With Crawford and others waving red flags, governments are trying to make software more accountable. Last December, the New York City Council passed a bill to set up a task force that will recommend how to publicly share information about algorithms and investigate them for bias. This year, France's president,

Emmanuel Macron, has said that the country will make all algorithms used by its government open. And in guidance issued this month, the UK government called for those working with data in the public sector to be transparent and accountable. Europe's General Data Protection Regulation (GDPR), which came into force at the end of May, is also expected to promote algorithmic accountability.

In the midst of such activity, scientists are confronting complex questions about what it means to make an algorithm fair. Researchers such as Vaithianathan, who work with public agencies to try to build responsible and effective software, must grapple with how automated tools might introduce bias or entrench existing inequity — especially if they are being inserted into an already discriminatory social system.

The questions that automated decision-making tools raise are not entirely new, notes Suresh Venkatasubramanian, a theoretical computer scientist at the University of Utah in Salt Lake City. Actuarial tools for assessing criminality or credit risk have been around for decades. But as large data sets and more-complex models become widespread, it is becoming harder to ignore their ethical implications, he says. "Computer scientists have no choice but to be engaged now. We can no longer just throw the algorithms over the fence and see what happens."

FAIRNESS TRADE-OFFS

When officials at the Department of Human Services in Allegheny County, where Pittsburgh is located, called in 2014 for proposals for an automated tool, they hadn't yet decided how to use it. But they knew they wanted to be open about the new system. "I'm very against using government money for black-box solutions where I can't tell my community what we're doing," says Erin Dalton, deputy director of the department's Office of Data Analysis, Research and Evaluation. The department has a centralized data warehouse, built in 1999, that contains a wealth of information about individuals — including on housing, mental health and criminal records. Vaithianathan's team put in an impressive bid to focus on child welfare, Dalton says.

The Allegheny Family Screening Tool (AFST) launched in August 2016. For each phone call to the hotline, call-centre employees see a score between 1 and 20 that is generated by the automated riskassessment system, with 20 corresponding to a case designated as highest risk. These are families for which the AFST predicts that children are most likely to be removed from their homes within two years, or to be referred to the county again because a caller has suspected abuse (the county is in the process of dropping this second metric, which does not seem to closely reflect the cases that require further investigation).

An independent researcher, Jeremy Goldhaber-Fiebert at Stanford



Police in Camden, New Jersey, use automated tools to help determine which areas need patrolling.

University in California, is still assessing the tool. But Dalton says preliminary results suggest that it is helping. The cases that call-centre staff refer to investigators seem to include more instances of legitimate concern, she says. Call screeners also seem to be making more consistent decisions about cases that have similar profiles. Still, their decisions don't necessarily agree with the algorithm's risk scores; the county is hoping to bring the two into closer alignment.

As the AFST was being deployed, Dalton wanted more help working out whether it might be biased. In 2016, she enlisted Alexandra Chouldechova, a statistician at Carnegie Mellon University in Pittsburgh, to analyse whether the software was discriminating against particular groups. Chouldechova had already been thinking about bias in algorithms and was about to weigh in on a case that has triggered substantial debate over the issue. In May that year, journalists at the news website ProPublica reported on commercial software used by judges in Broward County, Florida, that helps to decide whether a person charged with a crime should be released from jail before their trial. The journalists said

'If you want to be fair in one way, you might necessarily be unfair in another."

that the software was biased against black defendants. The tool, called COMPAS, generated scores designed to gauge the chance of a person committing another crime within two years if released.

The ProPublica team investigated COMPAS scores for thousands of defendants, which it had obtained through public-records requests. Comparing black and white defendants, the journalists found that a disproportionate number of black defendants were 'false positives': they were classified by COMPAS as high risk but subsequently not charged with another crime.

The developer of the algorithm, a Michigan-based company called Northpointe (now Equivant, of Canton, Ohio), argued that the tool was not biased. It said that COMPAS was equally good at predicting whether a white or black defendant classified as high risk would reoffend (an example of a concept called 'predictive parity'). Chouldechova soon showed that there was tension between Northpointe's and ProPublica's measures of fairness¹. Predictive parity, equal false-positive error rates, and equal false-negative error rates are all ways of being 'fair', but are

statistically impossible to reconcile if there are differences across two groups — such as the rates at which white and black people are being rearrested (see 'How to define 'fair"). "You can't have it all. If you want to be fair in one way, you might necessarily be unfair in another definition that also sounds reasonable," says Michael Veale, a researcher in responsible machine learning at University College London.

In fact, there are even more ways of defining fairness, mathematically speaking: at a conference this February, computer scientist Arvind Narayanan gave a talk entitled '21 fairness definitions and their politics' — and he noted that there were still others. Some researchers who have examined the ProPublica case, including Chouldechova, note that it's not clear that unequal error rates are indicative of bias. They instead reflect the fact that one group is more difficult to make predictions about than another, says Sharad Goel, a computer scientist at Stanford. "It turns out that that's more or less a statistical artefact."

For some, the ProPublica case highlights the fact that many agencies lack resources to ask for and properly assess algorithmic tools. "If anything, what it's showing us is that the government agency who hired Northpointe did not give them a well-defined definition to work with," says Rayid Ghani, who directs the Center for Data Science and Public Policy at the University of Chicago, Illinois. "I think that governments need to learn and get trained in how to ask for these systems, how to define the metrics they should be measuring and to make sure that the systems they are being given by vendors, consultants and researchers are actually fair."

Allegheny County's experience shows how difficult it is to navigate these questions. When Chouldechova, as requested, began digging through the Allegheny data in early 2017, she found that its tool also suffered similar statistical imbalances. The model had some "pretty undesirable properties", she says. The difference in error rates was much higher than expected across race and ethnicity groups. And, for reasons that are still not clear, white children that the algorithm scored as at highest risk of maltreatment were less likely to be removed from their homes than were black children given the highest risk scores². Allegheny and Vaithianathan's team are currently considering switching to a different model. That could help to reduce inequities, says Chouldechova.

Although statistical imbalances are a problem, a deeper dimension of unfairness lurks within algorithms — that they might reinforce societal injustices. For example, an algorithm such as COMPAS might purport to predict the chance of future criminal activity, but it can only rely on measurable proxies, such as being arrested. And variations in policing practices could mean that some communities are disproportionately targeted, with people being arrested for crimes that might be ignored in other communities. "Even if we are accurately predicting something, the thing we are accurately predicting might be the imposition of injustice," says David Robinson, a managing director at Upturn, a non-profit social-justice organization in Washington DC. Much would depend on the extent to which judges rely on such algorithms to make their decisions — about which little is known.

Allegheny's tool has come under criticism along similar lines. Writer and political scientist Virginia Eubanks has argued that, irrespective of whether the algorithm is accurate, it is acting on biased inputs, because black and biracial families are more likely to be reported to hotlines. Furthermore, because the model relies on public-services information in the Allegheny system — and because the families who used such services are generally poor — the algorithm unfairly penalizes poorer families by subjecting them to more scrutiny. Dalton acknowledges that the available data are a limitation, but she thinks the tool is needed. "The unfortunate societal issue of poverty does not negate our responsibility to improve our decision-making capacity for those children coming to our attention," the county said in a response to Eubanks, posted on the AFST website earlier this year.

TRANSPARENCY AND ITS LIMITS

Although some agencies build their own tools or use commercial software, academics are finding themselves in demand for work on public-sector algorithms. At the University of Chicago, Ghani has been working with a range of agencies, including the public-health department

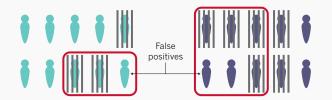
How to define 'fair'

Researchers studying bias in algorithms say there are many ways of defining fairness, which are sometimes contradictory.

Imagine that an algorithm for use in the criminal-justice system assigns scores to two groups (blue and purple) for their risk of being rearrested. Historical data indicate that the purple group has a higher rate of arrest, so the model would classify more people in the purple group as high risk (see figure, top). This could occur even if the model's developers try to avoid bias by not directly telling their model whether a person is blue or purple. That is because other data used as training inputs might correlate with being blue or purple.



A high-risk status cannot perfectly predict rearrest, but the algorithm's developers try to make the prediction equitable: for both groups, 'high risk' corresponds to a two-thirds chance of being rearrested within two years. (This kind of fairness is termed predictive parity.) Rates of future arrests might not follow past patterns. But in this simple example, assume that they do: as predicted, 3 out of 10 in the blue group and 6 out of 10 in the purple group (and two-thirds of those labelled high risk in each group) are indeed rearrested (indicated in grey bars in figure, bottom).



This algorithm has predictive parity. But there's a problem. In the blue group, 1 person out of 7 (14%) was misidentified as high risk; in the purple group, it was 2 people out of 4 (50%). So purple individuals are more likely to be 'false positives': misidentified as high risk.

As long as blue and purple group members are rearrested at different rates, then it will be difficult to achieve predictive parity and equal false-positive rates. And it is mathematically impossible to achieve this while also satisfying a third measure of fairness: equal false-negative rates (individuals who are identified as low risk but subsequently rearrested; in the example above, this happens to be equal, at 33%, for both purple and blue groups).

Some would see the higher false-positive rates for the purple group as discrimination. But other researchers argue that this is not necessarily clear evidence of bias in the algorithm. And there could be a deeper source for the imbalance: the purple group might have been unfairly targeted for arrest in the first place. In accurately predicting from past data that more people in the purple group will be rearrested, the algorithm could be recapitulating — and perhaps entrenching — a pre-existing societal bias. R.C.

of Chicago on a tool to predict which homes might harbour hazardous lead. In the United Kingdom, researchers at the University of Cambridge have worked with police in County Durham on a model that helps to identify who to refer to intervention programmes, as an alternative to prosecution. And Goel and his colleagues this year launched the Stanford Computational Policy Lab, which is conducting collaborations with government agencies, including the San Francisco District Attorney's office. Partnerships with outside researchers are crucial, says Maria McKee, an analyst at the district attorney's office. "We all have a sense of what is right and what is fair," she says. "But we often don't have the tools or the research to tell us exactly, mechanically, how to get there."

There is a large appetite for more transparency, along the lines adopted by Allegheny, which has engaged with stakeholders and opened its doors to journalists. Algorithms generally exacerbate problems when they are "closed loops that are not open for algorithmic auditing, for review,

or for public debate", says Crawford at the AI Now Institute. But it is not clear how best to make algorithms more open. Simply releasing all the parameters of a model won't provide much insight into how it works, says Ghani. Transparency can also conflict with efforts to protect privacy. And in some cases, disclosing too much information about how an algorithm works might allow people to game the system.

One big obstacle to accountability is that agencies often do not collect data on how the tools are used or their performance, says Goel. "A lot of times there's no transparency because there's nothing to share." The California legislature, for instance, has a draft bill that calls for risk-assessment tools to help reduce how often defendants must pay bail — a practice that has been criticized for penalizing lower-income defendants. Goel wants the bill to mandate that data are collected on instances when judges disagree with the tool and on specific details, including outcomes, of every case. "The goal is fundamentally to decrease incarceration while maintaining public safety," he says, "so we have to know — is that working?"

Crawford says that a range of 'due process' infrastructure will be needed to ensure that algorithms are made accountable. In April, the AI Now Institute outlined a framework³ for public agencies interested in responsible adoption of algorithmic decision-making tools; among other things, it called for soliciting community input and giving people the ability to appeal decisions made about them.

Many are hoping that laws could enforce such goals. There is some precedent, says Solon Barocas, a researcher who studies ethics and policy issues around artificial intelligence at Cornell University in Ithaca, New York. In the United States, some consumer-protection rules grant citizens an explanation when an unfavourable decision is made about their credit⁴. And in France, legislation that gives a right to explanation and the ability to dispute automated decisions can be found as early as the 1970s, says Veale.

The big test will be Europe's GDPR, which entered into force on 25 May. Some provisions — such as a right to meaningful information about the logic involved in cases of automated decision-making — seem to promote algorithmic accountability. But Brent Mittelstadt, a data ethicist at the Oxford Internet Institute, UK, says the GDPR might actually hamper it by creating a "legal minefield" for those who want to assess fairness. The best way to test whether an algorithm is biased along certain lines — for example, whether it favours one ethnicity over another requires knowing the relevant attributes about the people who go into the system. But the GDPR's restrictions on the use of such sensitive data are so severe and the penalties so high, Mittelstadt says, that companies in a position to evaluate algorithms might have little incentive to handle

the information. "It seems like that will be a limitation on our ability to assess fairness," he says. The scope of GDPR provisions that might give the public insight into algorithms and the ability to appeal is also in question. As written, some GDPR rules apply only to systems that are fully automated, which could exclude situations in which an algorithm affects a decision but a human is supposed to make the final call. The details, Mittelstadt says, should eventually be clarified in the courts.

AUDITING ALGORITHMS

Meanwhile, researchers are pushing ahead on strategies for detecting bias in algorithms that haven't been opened up for public scrutiny. Firms might be unwilling to discuss how they are working to address fairness, says Barocas, because it would mean admitting that there was a problem in the first place. Even if they do, their actions might ameliorate bias but not eliminate it, he says. "So any public statement about this will

> also inevitably be an acknowledgment that the problem persists." But in recent months, Microsoft and Facebook have both announced the development of tools to detect bias.

Some researchers, such as Christo Wilson, a computer scientist at Northeastern University in Boston, try to uncover bias in commercial algorithms from the outside. Wilson has created mock passengers who purport to be in search of Uber taxi rides, for example, and has uploaded dummy CVs to a jobs website to test for gender bias. Others are building software that they hope could be of general use in selfassessments. In May, Ghani and his colleagues released open-source software called Aequitas to help engineers, policymakers and analysts to audit machine-learning models for bias. And mathematician Cathy O'Neil, who has been vocal about the dangers of algorithmic decision-making, has launched a firm that is working privately with companies to audit their algorithms.

Some researchers are already calling for a step back, in criminal-justice applications and other areas, from a narrow focus on building algorithms that make forecasts. A tool might be good at predicting who will fail to appear

in court, for example. But it might be better to ask why people don't appear and, perhaps, to devise interventions, such as text reminders or transportation assistance, that might improve appearance rates. "What these tools often do is help us tinker around the edges, but what we need is wholesale change," says Vincent Southerland, a civil-rights lawyer and racial-justice advocate at New York University's law school. That said, the robust debate around algorithms, he says, "forces us all to ask and answer these really tough fundamental questions about the systems that we're working with and the ways in which they operate".

Vaithianathan, who is now in the process of extending her childabuse prediction model to Douglas and Larimer counties in Colorado, sees value in building better algorithms, even if the overarching system they are embedded in is flawed. That said, "algorithms can't be helicopter-dropped into these complex systems", she says: they must be implemented with the help of people who understand the wider context. But even the best efforts will face challenges, so in the absence of straight answers and perfect solutions, she says, transparency is the best policy. "I always say: if you can't be right, be honest."



Rhema Vaithianathan builds algorithms to help flag potential cases of child abuse.

Rachel Courtland is a science journalist based in New York City.

- Chouldechova, A. Preprint at https://arxiv.org/abs/1703.00056 (2017)
- Chouldechova, A., Putnam-Hornstein, E., Benavides-Prado, D., Fialko, O. & Vaithianathan, R. Proc. Machine Learn. Res. 81, 134-148 (2018).
- Reisman, D., Schultz, J., Crawford, K. & Whittaker, M. Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability (Al Now, 2018).
 Wachter, S., Mittelstadt, B. & Floridi, L. Sci. Robotics 2, eaan6080 (2017).

COMMENT

POLICY Conservation tries a new way to catalogue and synthesize research **p.364**

TOMMY TRENCHARD/PANOS



PUBLISHING Use ORCID and DOIs more for frictionless communication p.372

CONSERVATION Keep cruise ships off remote reef in South Pacific rich in species **p.372**



To improve vaccine uptake, nations can build on others' experience — if research is synthesized regularly and well.

Four principles for synthesizing evidence

Reward the creation of analyses for policymakers that are inclusive, rigorous, transparent and accessible, urge **Christl A. Donnelly** and colleagues.

o help address rising childhood obesity, researchers from Australia, Hong Kong and the United Kingdom collated and systematically analysed 55 studies, together involving tens of thousands of children. The result was one of the most influential medical reviews¹. It has been cited nearly 1,500 times since its publication in 2011, following nearly two years of work.

By contrast, it only took two days for the UK government to convene its Scientific Advisory Group for Emergencies (SAGE)following the 2011 disaster at the Fukushima nuclear plant in Japan, caused by an earthquake that hit the country's east coast (see go.nature.com/2jxotw6). Experts from within and outside government, including geologists, meteorologists, radiationhealth experts and behavioural scientists, rapidly modelled a range of possible scenarios. Within six days of the quake, they had advised that the risks to British nationals in Japan could be managed, and the UK government recommended that those outside the immediate exclusion zone stay put² (see also go.nature.com/2jptgnl).

These are both examples of evidence synthesis. This is the process of bringing together information and knowledge from many sources and disciplines to inform debates and decisions. Issues range from the impact of a pesticide on pollinators to who should be quarantined during a disease outbreak.

An accurate, concise and unbiased synthesis of the available evidence is arguably one of the most valuable contributions a research community can offer decision-makers. The common question 'What is the evidence?' could be usefully rephrased as

➤ 'Has sufficient synthesis of all the evidence been done in relation to that?'

Several organizations are already producing powerful examples of synthesized evidence. However, too few researchers and policymakers know about them; too few understand how to produce or commission good syntheses; and too many are reaching for information that is out of date, incomplete or biased, sometimes from just one study or researcher³. Even where good syntheses exist, they are often not available quickly enough: in the realm of public policy, it may be that a good-enough version available before a decision is made is much more valuable than a perfect version that arrives a day too late, provided the limitations imposed by doing it at speed are made clear.

Here we present a set of principles for good evidence synthesis for policy (see 'Four principles'). We are a group of academics, policymakers, evidence brokers and those responsible for research funding and publishing (including the editor-in-chief of this journal) in the United Kingdom — a world leader in science advice for policy.

We hope that these principles will make it easier for producers and users to commission, carry out, appraise, use and share high-quality evidence synthesis around the world.

WHY, HOW, WHAT, WHEN

Policy development is complex and frequently contested, and options can be viewed through several lenses. Evidence is an important lens, but not the only one. For example, stakeholders may have different personal and political values ('Do I morally object to culling badgers in order to tackle bovine tuberculosis?'), the objectives themselves may be disputed ('Is this about animal welfare or farm productivity or something else?') and there may be questions about the extent to which an 'ideal' solution can be delivered on the ground.

Given these multiple lenses, public debate and decision-making are best served by a clear, readily available synthesis of the current best evidence — which should stick to the lens of evidence alone if it is to be respected by policymakers.

Synthesis can take various shapes. Techniques range from a formal systematic review (as for the Cochrane Reviews common in medicine) to the rapid drawing together of evidence to inform an emergency situation (as for the Fukushima disaster or the 2014 Ebola epidemic in West Africa).

Formal systematic reviews follow a standard set of stages and can take many months to complete. They are the most established and comprehensive way to capture all the relevant evidence on a topic, and they can be used strategically to inform policy on topics that are predictable, enduring and recurrent — such as climate change or nutrition. But because this kind of study is time-consuming, important policy deadlines can be missed.

Rapid synthesis can respond more tactically to emergencies or, more commonly, to the day-to-day business of government. It can involve rapid evidence assessments, which are more targeted than a systematic review, with more-restricted search terms, evidence-gap maps (see, for example, go.nature.com/2tncfrq) and semi-structured interviews — techniques which ensure that more voices and views are considered and weighed, and which go beyond what a scientist would typically consider a 'review'.

Depending on its focus and purpose, synthesis may consider evidence of many kinds, including quantitative and qualitative data, published and unpublished academic literature, research conducted by industry or by non-governmental organizations, policy-evaluation studies from many countries and contexts, and expert and public opinion.

There are trade-offs between speed and thoroughness, of course, depending on priorities. But whatever the topic, time frame or methods, these four fundamental features should apply to every evidence synthesis.

INCLUSIVE

If policymakers are the target audience, they should be involved throughout — from

FOUR PRINCIPLES

These features help researchers, policymakers and others to commission, do, share, appraise and use evidence syntheses.

INCLUSIVE

- Involves policymakers and is relevant and useful to them.
- Considers many types and sources of evidence.
- Uses a range of skills and people.

RIGOROUS

- Uses the most comprehensive feasible body of evidence.
- Recognizes and minimizes bias.
- Is independently reviewed as part of a quality-assurance process.

TRANSPARENT

- Clearly describes the research question, methods, sources of evidence and quality-assurance process.
- Communicates complexities and areas of contention.
- Acknowledges assumptions, limitations and uncertainties, including any evidence gaps.
- Declares personal, political and organizational interests and manages any conflicts.

ACCESSIBLE

- Is written in plain language.
- Is available in a suitable time frame.
- Is freely available online.

designing the question to governing the process and interpreting the findings, although they should not mould that interpretation to support a particular policy. Policymakers might be less involved in the early stages if the aim is to scan the horizon for future priorities or to synthesize evidence on a topic that is yet to attract major policy interest, such as quantum computing.

Inclusivity helps to identify and make use of the full range of relevant evidence types, sources and expertise. During the Ebola epidemic, SAGE convened historians, anthropologists, behavioural scientists, engineers, mathematical modellers and infectiousdisease experts from around the world. The UK Government's Foresight projects typically involve around 200 scientists and scholars. Over 12 months or so, teams work with government departments, academics and experts from industry and elsewhere to identify where new or emerging science can inform long-term decision-making, on topics including flooding, cities and the future of the sea (see, for example, ref. 4).

RIGOROUS

Within the available time frame and resources, researchers should try to identify all the relevant evidence, before appraising its quality and analysing it. Synthesis which is not rigorous is bad science. It is also bad for policy, because policy informed by flawed science can lead to avoidable mistakes.

Rigorous synthesis always aims to minimize any bias that might distort the evidence or analysis. And personal prejudice has no place in evidence synthesis. Potential biases that cannot be avoided — for example, the fact that the literature on global agriculture comes predominantly from a small number of countries — must be disclosed and explained (see 'Transparent').

ee 'Transparent').
Cochrane (http://uk.cochrane.org) is an independent global network of researchers, professionals, carers and other people interested in health. It synthesizes evidence to inform health-care decisions made by national health services, funders, patients and others. The Campbell Collaboration (www.campbellcollaboration.org) provides a similar service for decision-making in education, social welfare, crime and justice, and international development, with reviews on topics including school start times, therapies for sexual offenders, and handwashing and sanitation behaviours in low- and middleincome countries. In both cases, co-ordinating groups manage the process in a way that minimizes bias — involving predefined methodologies, training for authors, peer review and often a significant amount of time. Producing a Cochrane or Campbell review can take more than two years.

Similarly, the Intergovernmental Panel on Climate Change (IPCC; www.ipcc.ch) ensures rigour in part by involving thousands



As technology and globalization alter how we work, synthesis reports help governments navigate change.

of authors and reviewers and spending five years or more crafting each assessment report. Multiple rounds of drafting aim to ensure that the synthesized evidence is as comprehensive and objective as possible.

These types of synthesis serve a specific strategic purpose. However, they need to be complemented by methods that can inform the rapid decision-making that goes on in governments. Oxford Martin Restatements, which began in 2013 and review the naturalscience evidence on policy issues ranging from bovine tuberculosis to ionizing radiation, provide one model for carrying out synthesis more quickly while maintaining rigour (see, for example, go.nature.com/2jrrmcm). They do this by involving 6–10 authors with different scientific points of view on a contested topic, and by seeking review comments from around 30-50 stakeholders before journal peer review.

TRANSPARENT

Evidence synthesis that is frank about its methods and limitations is likely to be more credible, replicable and useful — and can be better kept up to date.

The account of the study's methodology should include the search terms used, the databases and other evidence sources that were considered, when they were accessed, and the criteria that were used to determine which studies were included and why. Studies should explicitly acknowledge complexities and areas of consensus and contention, particularly where there are fundamental disagreements in the project team. This is important for evidence-based public debate, too. Outlining what is not known provides pointers to scientists, policymakers and funders on potential lines of enquiry to fill knowledge gaps.

The Restatements explicitly grade the strength of the evidence, and classify each paragraph using a set of descriptive codes. These include, at one end of the spectrum, "Data support a consensus based upon a

single well-powered study, or one or more pooled analyses with consistent results, or several lower-powered studies with consistent results," to, at the other end of the spectrum, "There is no consensus interpretation because the data are insufficient in quantity or too variable." The IPCC is also widely lauded for its clear assessments of the strength of evidence (qualified as 'limited', 'medium' or 'robust') and the degree of agreement among authors ('low', 'medium' or 'high').

ACCESSIBLE

Synthesized evidence will be quickly discarded by policymakers if they struggle to access it online or if the language is impenetrable. The full text and search terms should be published in an open-access repository to allow the synthesis to be extended, reproduced or updated in light of new evidence. In addition, reports should have a short summary in plain language.

A recent example is the Royal Society's 2017 report on machine learning⁵. It includes summaries for policymakers and for the wider public, with interactive graphics avail-

able online demonstrating machine learning in practice. The International Initiative for Impact Evaluation (3ie; www.3ieimpact. org), which promotes the use of evidence in devel-

"Synthesized evidence must be made available in time to contribute to the decisions it is intended to inform."

oping countries, typically produces three versions of each synthesis report. Two are targeted at policymakers (around 600 words) and practitioners (1,800 words), with a full review of 15,000 words in an academic journal. An example is a review of agricultural interventions for improved nutrition, published in *Global Food Security*⁶.

Timeliness is key. Synthesized evidence must be made available in time to contribute to the decisions it is intended to inform. In

the long run, habitually synthesizing evidence to provide answers to enduring questions — as is currently done to good effect in evidence-based medicine and for climate change — could reduce the need for more-rapid approaches.

NEXT STEPS

If done well, synthesis is a global public good. For some issues, evidence needs to be specific to the time and context in which decisions have to be made — who to vaccinate to halt a resurgence of polio on the border between Afghanistan and Pakistan, say, or whether to issue licences for fracking in the United Kingdom. But many issues — such as how to improve vaccine uptake — are common to decision-makers around the world. So syntheses that draw on evidence from different countries and contexts can have global value. Making synthesized evidence freely available for all means that knowledge can be shared and built on. Countries with a lower capacity to do research and to bring it together can benefit significantly — particularly if the process is collaborative and the evidence can be tested for local relevance and applicability.

The ultimate goal is to create an effective marketplace for synthesis in which policy-makers and commentators always seek the best evidence because they know it will be available, and researchers synthesize evidence because they know it will make a difference. Principles and exemplars provide a first step.

Moving towards this goal will require greater incentives and rewards for all stakeholders, through funding, evaluation, publishing and government practices, to promote work that adheres to the principles laid out here. To catalyse the necessary changes in the United Kingdom, the following organizations will promote and adopt the principles: the Royal Society, the Academy of Medical Sciences, UK Research and Innovation (UKRI), the Government Office for Science, the Department for Environment, Food and Rural Affairs, the Department of Health and Social Care and the UK Civil Service Policy Profession (an informal network for civil servants who work on government policymaking). Nature will also adopt all of the editorial principles but, given that it is a subscription journal, cannot currently commit to free access to all syntheses.

Ultimately, there needs to be a culture shift so that evidence synthesis is recognized as an exciting, intellectually challenging, high-status and respected activity for researchers, and one that underpins the identification of future research questions. For the United Kingdom, this might involve the evolution of the Research Excellence Framework (REF; www.ref.ac.uk), which will be updated in 2021. Such a shift could also encourage communities to work together continuously, allowing a mechanism for the refreshing of

COMMENT

synthesized evidence to be built in from the outset.

Several developments mean that the time is ripe. In the United Kingdom, the establishment of UKRI creates an opportunity to put in place mechanisms to support evidence synthesis as a complement to (and often a support for) primary research. Work by the academic community needs to be matched by an equal effort by policymakers to build science into policymaking systems. Several UK government departments have published Areas of Research Interest (ARIs; see, for example, ref. 7) — topics on which synthesized and new evidence would be most welcome. These are a valuable starting point for greater collaboration between departments and researchers. In addition, the Civil Service Policy Profession is developing a range of policymaking approaches to encourage the best use of evidence and to involve people from across a broad range of disciplines.

Internationally, there are numerous initiatives to improve the use of evidence in policymaking. The governments of Canada, New Zealand and the Australian state of New South Wales are adopting aspects of the What Works approach⁸, and there is growing interest in synthesis among groups such as Science Advice for Policy by European Academies (SAPEA) and the International Network for Government Science Advice (INGSA).

Synthesis requires brokerage at the interface of public life and academia. Collaboration will bring academics, policymakers, practitioners, funders and publishers closer to a world in which decision-making can be built on solid ground, not sand.

Christl A. Donnelly is professor of statistical epidemiology at Imperial College London, UK. Ian Boyd, Philip Campbell, Claire Craig, Patrick Vallance, Mark Walport, Christopher J. M. Whitty, Emma Woods, Chris Wormald.

e-mail: c.donnelly@imperial.ac.uk

- 1. Waters, E. et al. Cochrane Database Systematic Rev. CD001871 (2011).
- Grimes, R. W., Chamberlain, Y. & Oku, A. Sci. Diplomacy 3, 2 (2014).
- Sutherland, W. J. & Burgman, M. Nature 526, 317–318 (2015).
- 4. Foresight. *Future of the Sea* (UK Government Office for Science, 2018).
- Royal Society. Machine Learning: The Power and Promise of Computers that Learn by Example (Royal Society, 2017).
- Fiorella, K., L. Chen, R., Milner, E. & Fernald, L. Global Food Security 8, 39–47 (2016).
- UK Department for Environment, Food and Rural Affairs. Defra Group Areas of Research Interest (DEFRA, 2017).
- 8. UK Government. The What Works Network: Five Years On (UK Government, 2018).

Supplementary information accompanies this article online: see go.nature.com/2l46tuq.



Storks in Malpartida de Cáceres, western Spain, nest on purpose-built poles in a conservation area.

A fresh approach to evidence synthesis

Systematic reviews have transformed medicine. For fields in which data are sparse and patchy, a more cost-effective means of appraisal is needed, argue William J. Sutherland and Claire F. R. Wordley.

In 1990, researchers conducted a systematic review of studies investigating the use of corticosteroids in women who were at risk of giving birth prematurely¹. (The steroids were administered to reduce the chances of the women's pre-term babies experiencing respiratory issues.) The results of the first

of these clinical trials, published in 1972, had indicated that the treatment worked². But it was not widely adopted, because of concerns about potential side effects and the quality of the evidence. Indeed, the effectiveness of corticosteroids was conclusively established only with the 1990 review. Tens of thousands of lives have



probably been saved since then because of this intervention.

Among the countless methods regularly used to pull together evidence from different sources (see page 361), only systematic reviews critically appraise findings in a comprehensive manner. Reviews such as the corticosteroid one, which can include meta-analysis (whereby data from multiple studies are pooled and analysed together), have transformed medicine. Moreover, the improvements to medical practice have inspired the use of systematic reviews in other fields, such as policing³.

Yet such reviews are enormously time-consuming and expensive. We think that in fields in which data are sparse or patchily distributed, or where studies vary greatly in design and generalizability — as is the case in biodiversity conservation, international development and education, for example — a different approach might often be more appropriate. In our view, given the expansion of available studies, it has never been more important to have a large-scale, cost-effective way of rigorously appraising information for applied fields.

To address this need, we have a developed

a method that we call subject-wide evidence synthesis. Here, we lay out how it works.

SEARCHABLE SYNOPSES

Provided enough studies exist, systematic reviews and meta-analyses are invaluable for providing clear answers to focused questions. But it is hard to conduct a meta-analysis if only a few studies exist, or when those that do exist use different methods or measure different variables. Furthermore, such analyses are labour-intensive and expensive: in medical fields, systematic reviews generally take about a year to conduct and can cost between US\$30,000 and \$300,000 each⁴.

Another approach, called systematic mapping, is more broad-brush. But this typically does not describe the findings of the research, and so cannot be used to answer questions about policy (see 'Review or map?').

The approach we're advocating — subject-wide evidence synthesis — combines elements of systematic reviewing and mapping, along with other techniques, to provide an industrial-scale, cost-effective way to synthesize evidence. It is not intended to replace the use of systematic reviews, but it does provide a rigorous way to synthesize information when data are unevenly or thinly distributed, or highly variable in focus.

We have developed this approach in a project called Conservation Evidence (www. conservationevidence.com), which aims to assess the impact of conservation interventions for all species and habitats worldwide⁵.

After identifying broad subject areas such as bird conservation or reptile conservation, we established which interventions are likely to be relevant, and defined the criteria for including papers. For this, we drew on the expertise of advisory panels; for instance,

16 specialists in bird conservation helped us to draw up a list of 455 conservation interventions relevant to birds — from the use of model birds to lure species towards a safe

"Our team of researchers has searched every issue of nearly 250 journals for tests of some conservation intervention."

location for nesting, to the use of signs and access restrictions to protect nesting birds from human disturbance.

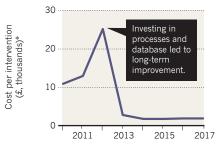
Collating the required information involves manually searching every paper in every issue of the journals we deemed relevant. Many papers can be excluded from our database simply by reading their title. For others, it's necessary to read the abstract, or even the entire paper, before deciding whether it should be extracted, tagged and stored.

For each intervention, a paragraph summarizes the key findings of all the studies that have been conducted; each study is

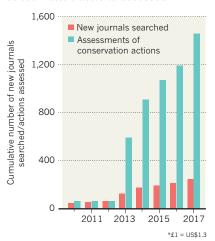
COST-EFFECTIVE

Manually searching journals to find studies on actions designed to conserve species is costly at first. But as the searches accumulate, subsequent evidence syntheses require fewer resources.

Spending per intervention



New journals searched and conservation actions accessed



summarized in an additional paragraph that provides information about the design, sample sizes and location. An expert assessment of the likely effectiveness of the intervention is also provided. All this information is collated in a 'synopsis'; for bird conservation, the current print version of the synopsis is a 466-page book.

The idea is to provide users with intervention assessments that they can use to pursue both broad and narrow questions. For example, conservationists might ask, 'How can we reduce fulmar by-catch at sea?' They could look at the assessments of five interventions designed to reduce seabird by-catch, each of which has been tested on species including fulmars. Or they could look at the assessments of 22 interventions that aim to reduce seabird by-catch but that haven't all been tested on fulmars specifically. Alternatively, a practitioner asking, 'What can be done to conserve seabirds?' might want to read about all 48 interventions pertaining to the conservation of seabirds.

Since 2004, our international team of researchers has searched every issue of nearly 250 journals for tests of some conservation intervention. And we've

conducted more than 1,700 assessments of interventions⁵ (see 'Cost-effective').

SHARPENING THE FOCUS

For most conservation interventions, insufficient evidence exists for investigators to be able to conduct a meta-analysis for each species or genus. However, by being apprised of studies that examine how a particular intervention has worked for an order — for birds in general, say — practitioners can better weigh up the chances of success for their intended programme.

Because subject-wide evidence synthesis entails scanning all the papers from every issue of the journals selected, it can unearth unusual interventions that would not necessarily have been identified on the basis of predetermined criteria for paper inclusion. For example, in our searches, we came across a study in which researchers had added snakeskins to nest boxes to deter mammal predators⁶. This was not on our original list of interventions.

Moreover, in subject-wide evidence synthesis, all papers relevant to the broader discipline (in this case, biodiversity conservation) are extracted, tagged and stored when searching a journal. This means that when the next synopsis is written (on amphibian conservation, say), those producing it just need to add the specialist amphibian journals to the 'bank' of journals that have already been searched.

Because the literature on an entire subject area has already been searched and summarized, focused topics can be investigated more nimbly. Also, subject-wide evidence syntheses can be easily updated, because the format for reporting results is standardized. Every new edition of each journal can be searched, a summary paragraph uploaded for each new paper, and the key messages concerning specific interventions changed to reflect the latest findings. For the Conservation Evidence project, the ideal would be to update every synopsis every second year. We are currently updating the first editions for birds and bats.

Lastly, subject-wide evidence synthesis can provide a starting point for different kinds of review. For instance, it identifies areas that are rich in evidence and thus suitable for systematic reviews.

Ultimately, subject-wide evidence synthesis should result in a resource that is concise, easy to navigate and comprehensible to non-scientists. In 2017, the Conservation Evidence website had 15,000-25,000 page views each month.

By using this method instead of search terms, some papers in obscure journals might be missed. At Conservation Evidence, we aim to continually expand our range of searched journals to reduce this. We are also trying to include relevant 'grey', or unpublished, literature in our searches, for instance by asking organizations such as Scottish Natural Heritage to share reports. Another concern is that evaluations of interventions might be biased, depending on the expertise of the assessors. We try to minimize this by using multiple anonymous rounds of scoring and a large team of assessors (the Delphi technique)⁷. And we provide median instead of mean scores, because medians are less influenced by outliers.

Collating data on such a scale is expensive; our bird-conservation synopsis, the result of searching 35 journals, has cost nearly £350,000 (US\$467,000). But once the papers have been extracted, many synopses can be produced. Indeed, the costs of producing assessments of interventions and synopses decline over time as each investigator builds on the efforts of others (see 'Cost-effective'). Overall, this approach is much more costeffective than standard systematic reviews that rely on the use of search terms.

A MULTIPURPOSE TOOL

As the scientific literature continues to grow, locating and collating new papers in evidence syntheses is becoming increasingly challenging8. Advances in artificial intelligence and machine learning could make it easier to perform tasks such as locating papers for defined topics using search

"In 2017, the Conservation **Evidence** website had 15.000-25.000 page views each month."

terms, categorizing papers as relevant for further consideration, and producing systematic maps⁹. But for all fields, assessing the quality of individual studies. writing up summa-

ries and so on will continue to require skilled humans, at least for the foreseeable future.

Because the cost-effectiveness of subjectwide evidence synthesis kicks in only when a large enough evidence bank has been developed, long-term funding to develop, sustain and update subject-wide evidence synthesis projects will be crucial.

So far, Conservation Evidence has been supported mainly by philanthropists, alongside research councils, industry and the UK government. Other projects might require core government funding, as has been provided to the UK What Works centres, which help to ensure that high-quality evidence shapes public-sector decision-making.

Our hope is that subject-wide evidence synthesis will prove as useful in disciplines such as international development as it seems to be in conservation. Ultimately, the proven usefulness of this approach in a range of fields will persuade practitioners that it is an indispensable part of the toolkit when it comes to collating knowledge to inform policy decisions. ■

William J. Sutherland is professor of conservation biology at the University of Cambridge, UK, and the founder and director of Conservation Evidence. Claire F. R. Wordley is a postdoctoral research associate on the Conservation Evidence project at the University of Cambridge. e-mails: w.sutherland@zoo.cam.ac.uk; cfw41@cam.ac.uk

- **REVIEW OR MAP?**
- Two well-used ways to pool information

In systematic reviews, investigators generally pose a focused question, such as: 'Is surgery an effective treatment for knee osteoarthritis?' They then write an a priori protocol paper to lay out their criteria for including studies, and explain how they will conduct their analysis before carrying out the review itself.

For the actual review, a PubMed or Scopus search for the terms 'knee surgery' or 'knee osteoarthritis', say, might give several thousand hits. By reading article titles and abstracts, researchers pare down their selection to those studies

that meet the predetermined criteria, and then conduct a qualitative analysis or meta-analysis on those.

In systematic maps, search terms are used to address open-framed questions, such as how many studies have been conducted on a particular topic, or how those studies have been conducted. This technique, which also often involves the publication of a priori methods, is used in fields from education to sustainability. In particular, it can identify knowledge gaps and hot spots for research, and so indicate priorities for future efforts 10,11.

- 1. Crowley, P., Chalmers, I. & Keirse, M. J. BJOG Int. J. Obstet. Gynaecol. 97, 11-25 (1990).
- Liggins, G. C. & Howie, R. N. Pediatrics 50, 515-525 (1972).
- 3. Sherman, L. W. Evidence-based Policing (Police
- Foundation, 1998). Dicks, L. V., Walsh, J. C. & Sutherland, W. J. *Trends* Ecol. Evol. 29, 607-613 (2014).
- Sutherland, W. J., Dicks, L. V., Ockendon, N., Petrovan, S. O. & Smith, R. K. (eds) What Works in Conservation (Open Book, 2018).
- Medlin, E. C. & Risch, T. S. Condor 108, 963-965 (2006).
- 7. Mukherjee, N. et al. Methods Ecol. Evol. 6, 1097-1109 (2015).
- Westgate, M. J. et al. Nature Ecol. Evol. 2, 588-590 (2018).
- O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. & Ananiadou, S. Syst. Rev. 4, 5 (2015). 10. James, K. L., Randall, N. P. & Haddaway, N. R. Environ. Evid. 5, 7 (2016).
- 11.McKinnon, M. C. Nature 528, 185-187 (2015).



North Korean leader Kim Jong-un watches the launch of a ballistic missile. The country flew the same type of missile over Hokkaido, Japan, in September 2017.

Unblock the path to peace in North Korea

Researchers must improve ways of verifying nuclear disarmament and collaborate on nuclear power, argues **R. Scott Kemp**.

here's a lesson to take from the speed with which the US–North Korean dialogue has oscillated between thinly veiled threats of nuclear annihilation and the promise of cooperation: that constructive dialogue between nations is never more than a political decision away. Openings can come quickly; when they do, it pays to be prepared.

Unfortunately, there are major technical gaps in the ability of science to verify the eventual nuclear disarmament of North Korea. Closing these gaps will be crucial to moving from the generalities produced for the global stage to specifics on the ground. Given that solutions will take time to develop, scientists and engineers need to focus on these challenges — and soon.

Understandably, many experts remain

sceptical that diplomacy alone will deliver a lasting solution. All previous attempts to work with North Korea have floundered because of deep-rooted mistrust and reticent compliance from all sides. The same could happen again, especially if expectations grow too quickly. But there are also solid reasons for optimism. The motivations of all parties to stay the course are higher than ever: North Korea is on the verge of producing a significant nuclear arsenal, making the consequences of failure extremely high. On the positive side, the benefits of successful diplomacy could bring economic gains for both sides. North Korea holds an estimated US\$10 trillion in untapped mineral reserves, including what might be the world's largest single deposit of rare-earth metals. Structural changes that

have been unfolding inside North Korea for the past six years suggest that now is the time to begin a measured engagement (see 'Reasons for optimism').

This is not to say that the world should expect steady progress. The Iran nuclear agreement, by comparison, took 13 years to bring about. I was one of the scientists working to make that happen, and I watched first-hand as planners experienced setback after setback. Similarly, during the cold war, arms control between the United States and Soviet Union was achieved through punctuated evolution rather than through a single grand agreement.

In several ways, the former Soviet Union might be a better model for understanding North Korea than, say, Libya or Iran. North Korea can use its missiles and nuclear

weapons for leverage; it is motivated by a complex mix of economic distress and security fears; and it holds an almost pathological distrust of the United States.

Thus, as with the Soviet Union, the world should expect North Korea to hedge its bets, engage in brinkmanship and cheat on its agreements. An enlightened approach can nevertheless be resilient. For example, the United States stood by the 1972 Anti-Ballistic Missile Treaty in 1985 when the administration of then US President Ronald Reagan learnt that the Soviet Union had built prohibited radars. Instead of tearing up the treaty, Reagan used the evidence to pressure the Soviets into accepting a subsequent agreement that placed more restrictions on their arsenal.

One crucial difference is that denuclearization — total nuclear disarmament and the removal of supporting nuclear facilities — was never on the table with the Soviet Union as it is with North Korea. As a consequence, many of the verification challenges associated with denuclearization were never solved.

South Africa is the only case in history that comes close: it gave up its nuclear weapons voluntarily in 1991 before the apartheid-era National Party ceded power to the African National Congress. The outgoing government did not want to bequeath nuclear weapons to its successors. The new government cooperated by giving international inspectors access to sites. Technical

discrepancies still cropped up¹. The world ultimately accepted South Africa as being free from nuclear weapons for political, rather than technical, reasons. North Korea will be a very different case.

KNOWN UNKNOWNS

At present, at least three challenges would impede successful verification².

First, there is no known way of detecting secret centrifuge facilities for enriching uranium. These can be smaller than an office building, would easily fit into an underground tunnel, consume very little electricity and release almost no thermal or chemical emissions — all of which makes them hard to find. Based on the speed with which North Korea built the one enrichment facility it has made public, there is good reason to believe that it operates at least one other research facility. It has probably built other full-size production facilities, too. These facilities need to be found and dismantled.

Second, there is currently no way of estimating how much weapons-grade uranium North Korea has produced. A forensic method is needed to reconstruct the production history for each of the uranium-enrichment facilities that might be discovered. This process will help to confirm the completeness of declarations and ensure that North Korea is not keeping a secret cache of enriched uranium for future use.

REASONS FOR OPTIMISM

Kim elevates the economy over military might

There is growing evidence that North Korea is undergoing a real shift. Its leader, Kim Jong-un, stepped into his father's shoes 6.5 years ago, at age 28. He began his reign with displays of power, purging internal challengers and fast-tracking the development of nuclear weapons and long-range missiles. It was a discouraging start, but after cementing his authority over a sceptical military, he began to plant seeds of change.

In March 2013, Kim unveiled a new party line, byungjin — referring to 'parallel development' of the economy and nuclear weapons. Although the basis of the policy is not peaceful, it was nonetheless a liberalizing segue from songun, the 'military first' policy of his father. He elevated his economic advisers to the front of his retinue, where military officials used to stand. He has also elevated the party in terms of its relationship to the military.

At the 7th Congress of the Workers' Party of Korea in May 2016 — the first such congress in 36 years — he adopted a policy enshrining scientific research and technological innovation, not military discipline, as the motive force for national progress. This April, Kim announced that his transitional *byungjin* policy had been fulfilled; the new party line is now to devote all available resources to economic development.

On the ground, there is evidence of a new urban-services economy. The number of official shops in North Korea where ordinary citizens can buy basic goods has nearly doubled since 2010. Markets are now more formally organized, with merchants renting stalls from the government. Several private taxi companies now compete for business in Pyongyang. And the government is developing tourism sites in areas of unspoiled beauty to welcome outsiders and their cash.

The Western-educated North Korean leader might understand that his success at home, and the survival of his nation, depends on finding a controlled way to engage the outside world. R.S.K.

Third, the international community does not know how much uranium or plutonium was consumed in North Korea's weapons tests. Weapon designs can use a wide range of quantities, and only a small fraction is destroyed during the explosion. As such, on the basis of the explosive yield, it is not possible to estimate how much uranium or plutonium fuel was removed from stockpiles. North Korea could claim that large quantities of uranium or plutonium were used to build its early weapons, while actually holding significant amounts in reserve.

In the absence of technical tools, inspectors will try their best using interviews, logbooks and other records — but these could easily be incomplete or forged. Robust technical methods would provide an extremely valuable layer of confidence that does not rely as heavily on trust.

SCIENCE PRIORITIES

Solutions to these long-standing challenges lie at the frontiers of research. For example, if chemists or materials scientists could develop a way to sense uranium–fluorine bonds in aerosols at levels between parts per billion and parts per trillion, this could pave the way for exposing clandestine uranium-enrichment plants. Detecting uranium or fluorine alone would not work, because both are relatively abundant, but the uranium–fluorine bond is unique to nuclear programmes.

Analytical methods might also be developed to check declarations about the past production of enriched uranium. There are ways to examine chemical residues in facilities. None are reliable enough, because they are sensitive to operating conditions, such as humidity, that can vary widely from plant to plant and over time. Methods that do not rely on chemistry, or that can combine multiple lines of information to control for variations, might help to solve this challenge.

The third problem of estimating the amount of material used in past weapon tests might be conquered by geochemists, hydrologists and petroleum engineers. They could drill into the mountain where North Korea performed its tests, and assess the amount of plutonium or uranium associated with each detonation. Given the disturbed rock zones, the oxidizing environment and water transport of isotopes through the mountain, this presents a challenging, long-term study.

The verification requirements outlined here are well suited to government laboratories, but history suggests that university and industry scientists can play a valuable part in forging international agreements as well. For example, in 1988, a group of non-government scientists from both the United States and the Soviet Union shored up confidence



Journalists from outside North Korea record the destruction of part of the country's nuclear test site at Punggye-ri.

in the proposed Comprehensive Test Ban Treaty by demonstrating that even small, conventional explosions could be detected at a Soviet nuclear test site, using seismic monitoring. In 1991, physicist Thomas Neff, then at the Massachusetts Institute of Technology in Cambridge, proposed turning the uranium from Soviet nuclear weapons into fuel for civil nuclear reactors. For almost 20 years, starting in 1995, about 10% of US electricity was generated from dismantled Soviet warheads.

Academics are often more able than government scientists to propose novel ideas. They can proceed independently of political posturing, and they can communicate to the public or to key decision-makers without layers of political filtering³.

Verification will be slow (as will diplomacy). It is unlikely that we will be able to achieve high confidence in the accuracy and completeness of North Korean declarations in the next decade.

In the meantime, engaging with North Korean scientists and engineers will itself build confidence. During the cold war, the Academy of Sciences of the USSR undertook a study of how it could ever trust the exceptionally adroit United States. The study concluded, "The criterion [for] a verification regime is the reduction to a minimum of pretexts for mutual suspicion about possible and significant concealed

violations ... even if it fails to detect a violation right away." The United States commissioned the JASON group of science advisers to do a similar study. It concluded "verification will of necessity be less than perfect ... it must rely on difficult political/ strategic judgments, as well as technical ones." In short, trust comes as much from the people involved and the relationships they forge as it does from the data.

There are opportunities to begin building useful relation-

ships now. One way is to use civil nuclear energy as a bridge to moresensitive nuclear issues, because it involves many

"Engaging with North Korean scientists and engineers will itself build confidence."

of the same technologies, experts and facilities. Earlier this year, North Korea completed major construction on what it calls an experimental light-water power reactor, at Yongbyon. However, the country does not have the capability to test fuel for the reactor to international safety standards. South Korea has both the test facilities and the decades of experience needed. Cooperation between the two nations in this area would help to establish trust, and would reduce the risk of an accident that could shower the citizens of South Korea with radioactive fallout.

The full closure of North Korea's nuclear-weapons programme will take many years of political horse-trading, inspections and verification work. There will be periods of optimism and of tension. When the politics do align, government advisers will be scrambling for solutions to the hard technical problems that impede progress. In anticipation of those moments, researchers who wish to unblock the path to peace should begin working now.

R. Scott Kemp is associate professor of nuclear science and engineering at the Massachusetts Institute of Technology (MIT), and director of the MIT Laboratory for Nuclear Security and Policy in Cambridge. In 2010 and 2011, he served as science adviser in the US State Department's Office of the Special Advisor for Nonproliferation and Arms Control. e-mail: rsk@mit.edu

- von Baeckmann, A., Dillon, G. & Perricos, D. IAEA Bull. 37-1, 42–48 (1995).
- Kemp, R. S. Ann. Rev. Earth Planet. Sci. 44, 17–35 (2016).
- 3. Von Hippel, F. Citizen Scientist (American Institute of Physics, 1991).
- Vasiliev, A., Gerasev, S. & Oznobischev, O. Sea-Launched Cruise Missiles: Verification Solutions (Institute for US and Canadian Studies, 1988).
- Drell, S. et al. Verification of Dismantlement of Nuclear Warheads and Controls on Nuclear Materials (JASON, The MITRE Corporation, 1993)

ASTRONOMY

Maria Mitchell at 200

Richard Holmes celebrates the pedagogic fire and salty opinions of the pioneering astronomer and advocate of women's rights on the bicentenary of her birth.

The was the first woman in the United States to become a professional astronomer, and a dauntless champion of science education for women. Maria Mitchell, whose bicentenary is celebrated this August, was a scientific revolutionary. That is encapsulated in her prophetic speech, 'The Need for Women in Science', delivered in 1876 to the Fourth Congress of the Association for the Advancement of Women, in Philadelphia, Pennsylvania. It posed a historic challenge. Mitchell declared that the laws of nature are discovered not through "the hurry and worry of daily toil; they are diligently sought ... And until able women have given their lives to investigation, it is idle to discuss the question of their capacity for original work." Or, as she put it in her journals: "better to be peering in the spectrograph than on the pattern of a dress".

Mitchell's agile mind, pedagogic fire and salty opinions bring extraordinary animation to her varied collection of scientific papers, articles, notebooks and journals. Some were first published by her sister Phebe Mitchell Kendall, in the 1896 Maria Mitchell: Life, Letters, and Journals, just seven years after her death. More recent volumes include Henry Albers's edited Maria Mitchell: A Life in Journals and Letters (2001) and Renée Bergland's Maria Mitchell and the Sexing of Science (2008). Independent, combative and original, Mitchell became a major public figure by the end of her life. She "stands out clear and conspicuous", noted her Scientific American obituary in July 1889, "like an evening star in the heavens she loved so well to study".

She was born in 1818, into a large Quaker family on the island whaling station of Nantucket, off the Massachusetts coast. It was a place of horizon-gazers, seafarers and lighthouse keepers, where men were often away and Quakerism honoured gender equality. From early childhood she was encouraged to pursue science by her beloved father, a director of the local bank and an amateur astronomer with contacts at Harvard University's observatory in Cambridge, Massachusetts. Together they scanned the skies using a Dollond telescope from the rooftop 'walk' of their house. Early on, Mitchell revealed extraordinary observational powers, natural mathematical gifts and unusual sensitivity to stellar movements and colours.

At 17, she opened her own school; a year



Maria Mitchell (left) and Mary Whitney in the observatory at Vassar College.

later, she was appointed supervisor of the local library, the Nantucket Atheneum. Her constant companion was a notebook, carried in a capacious pocket. Her speech was direct, her ideas increasingly radical. "We cannot accept anything as granted," she wrote in her journal, "beyond the first mathematical formulae. Question everything else."

Like the German astronomer Caroline Herschel (1750–1848), Mitchell made her name through discovering a comet. On 1 October 1847, on the roof of the Pacific Bank, she spied a new "telescopic comet" five degrees above the pole star. She published a preliminary notice of her finding in the journal of Britain's Royal Astronomical Society on 12 November, giving her a claim to the gold medal established by the Danish King Frederick VI for the first sighting of any new comet, detectable only by telescope, anywhere in the world.

Mitchell's claim to priority was authentic,

but closely matched by astronomers in Italy, Germany and Britain. She found an enthusiastic champion in Edward Everett, president of Harvard University, who argued her case in various scientific journals, and even wrote personally to the Danish consul in Washington DC. He also wrote drily to a friend in late 1847: "It would be pleasant to have the Nantucket girl carry off the prize from all the greybeards and observatories in Europe."

She did, and it became known as Miss Mitchell's Comet. Recognition swiftly followed. By the age of 32, she had become both the first woman elected to the American Academy of Arts and Sciences, and the first to enter the American Association for the Advancement of Science. In 1849, she moved with her father to Boston, Massachusetts, and was appointed an official 'computer' calculating tables for the US *Nautical Almanac* (see S. Nelson *Nature* 539, 491–492; 2016). She was assigned the orbit of Venus.

That orbit was much too restricting. In 1857, she set out alone to visit the great European observatories, from Greenwich, UK, to Berlin. She took with her one of the earliest star photographs from Harvard, and a volume of Lord Byron's poetry. Her journals record encounters with George Airy, Britain's astronomer royal ("the Bear of Blackheath"); William Whewell, master of Trinity College, Cambridge, and coiner of the term 'scientist'; astronomer John Herschel (Caroline's nephew, and "a better listener than any man I have met in England"); and the great explorer Alexander von Humboldt.

Not all these meetings went smoothly. She was "riled" by Whewell's chauvinist teasing while dining at Trinity. He mocked the "disgusting" poems of Elizabeth Barrett Browning, and scorned ideas of extraterrestrial intelligence, as contrary to the design of a benevolent creator. In her travel journal, Mitchell noted that Whewell's 1853 book *The Plurality of Worlds* "reasons to this end: "The planets were created for this world; this world for man; man for England; England for Cambridge; and Cambridge for Dr Whewell!"

But her encounter in Florence with British mathematician and writer Mary Somerville was a true meeting of minds. Then 77, Somerville "came tripping into the room, speaking at once with the vivacity of a young person". Mitchell was dazzled by Somerville's encyclopaedic enthusiasms, spanning recent discoveries "in chemistry or the discovery of gold in California, of the nebulae ... of the planets". Above all, Somerville's books — notably the 1834 On the Connexion of the Physical Sciences (see R. Holmes Nature 514, 432-433; 2014) — had expanded Mitchell's sense of the scientific world, and confirmed her refusal to be defined by gender. In Rome, Mitchell discovered — as Somerville had, years earlier — that the Vatican Observatory was closed to women. Eventually, Mitchell secured



Maria Mitchell argued for women's education.

permission to visit, but only during the day.

Mitchell later wrote a study of the friendship struck between astronomer Galileo Galilei and the English poet John Milton in the 1630s. She loved Milton's evocation in *Paradise Lost* of Galileo near Florence, entranced by the Moon, "whose orb/ Through optic glass the Tuscan artist views/ At evening, from the top of Fesole".

In 1865, she was appointed professor of astronomy at the newly founded Vassar College in Poughkeepsie, New York — one of the earliest woman-only US institutes of higher

"She railed against the era's domestic drudgery, urging her students to pursue full-time careers in science."

learning. Her yearly salary was US\$800, a figure that Mitchell vigorously disputed after discovering it was a fraction of that paid to male professors. The Vassar observatory, designed

by mathematician Charles Farrar, had an 8.4-metre dome revolved by 16 cast-iron pulleys, along with transit and chronograph rooms; it was regarded as the best equipped in the United States, after Harvard's. Mitchell lived at the observatory for the next 23 years, and was supplied with a superb 30-centimetre reflector telescope built especially for her use by instrument-maker Henry Fitz.

Under a bust of Somerville, Mitchell mentored a brilliant circle of devoted female students. Her first intake included Mary Whitney, future astronomer, supporter of women's rights and Vassar professor, who was part of an inner group of six young stars known as the Hexagon. True to Mitchell's esteem for the arts, they started a tradition of annual 'Dome parties', mixing poetry and astronomy with strawberries and cream. The imagination, Mitchell asserted, is part of science, which is "not all mathematics, or all

logic, but is somewhat beauty and poetry".

As a lecturer, Mitchell became known for pithy sayings. One was: "Study as if you were going to live forever; live as if you were going to die tomorrow." On Isaac Newton's genius: "Newton rolled up the cover of a book; he put a small glass at one end, and a large brain at the other — it was enough." She explained the use of spectroscopy: "The Astronomer breaks up the starlight just as the geologist breaks up the rock with his hammer, and with similar results, he finds copper, sodium and other elements in sun and stars."

She felt that women had a gift for observational astronomy: "The eye that directs a needle in the delicate meshes of embroidery will equally well bisect a star with the spider web of the micrometer." She railed against the era's domestic drudgery, urging her students to pursue full-time careers in science. Above all, she argued that higher education alone would give women independence of mind and drive: "Until women throw off this reverence for [male] authority they will not develop. When they do this, when they come to truth through their investigations ... their minds will work on and on, unfettered."

Mitchell's fieldwork was intrepid. In 1873 she travelled to an observatory in Russia, just outside St Petersburg. In 1878, she led an all-female expedition of her best students to observe a solar eclipse from Denver, Colorado (see J. Pasachoff Nature 545, 409-410; 2017). They travelled 3,000 kilometres on the newly completed transcontinental Pacific Railroad, camping and setting up their telescopes on an open plain with a view of the Rocky Mountains. During her final months at Vassar, Mitchell wrote a moving journal entry, comparing the quiet continuity of her meticulous observations of annular eclipses across a gap of more than 50 years. In 1831, she and her father gazed from their Nantucket rooftop; in 1885, she was surrounded by students in a beautifully appointed observatory. "Both days were perfectly cold and clear." Retirement in 1888 did not suit Mitchell. She died within a year, with a telescope still at her window.

This is a legacy to reckon with. Her archives are treasured at Vassar, and a museum and an association in her name flourish in Nantucket. Her beautiful Henry Fitz telescope has gleaming pride of place in the National Museum of American History, Washington DC. Above all, she should be remembered for her inspirational science teaching, the passionate ex-Quaker and bold proto-feminist so vividly combined. One of her students recalled: "A chance meeting with Miss Mitchell ... gave one always an electric shock. At the slightest contact, a spark flashed." We can catch it still. ■

Richard Holmes is the author of The Age of Wonder and This Long Pursuit. e-mail: richard.holmes.biog@gmail.com

Correspondence

Risks of tech metals under surveillance

Our understanding of the possible environmental and ecological toxic effects of technology-critical elements after extraction is not as limited as Winfred Espejo and colleagues imply (*Nature* 557, 492; 2018). Research is being done into possible risks, especially in Europe, and the community of scientists involved is growing.

For instance, a European Cooperation in Science & Technology (COST) Action has been evaluating these risks since 2015. As well as organizing training schools and advanced workshops, the TD1407 action facilitates network and capacity building (see www.costnotice. net).

Many technology-critical elements have not been investigated (M. Filella and J. C. Rodríguez-Murillo et al. Chemosphere 182, 605-616; 2017), but some have attracted attention. An example is gadolinium, mainly because of its rapid accumulation in surface waters. Several others cited by the correspondents have long been under investigation because they have applications that pre-date their current use in technology. Platinum is one such example. Montserrat Filella University of Geneva, Switzerland. Ishai Dror Weizmann Institute of Science, Rehovot, Israel. **Sebastien Rauch** Chalmers University of Technology, Gothenburg, Sweden. montserrat.filella@unige.ch

Replication drive for humanities

Research in humanities disciplines such as anthropology, archaeology, linguistics and theology can learn from replication failures in the biomedical and social sciences (go.nature.com/2stme7r).

Replication studies are not unprecedented in the humanities. The deciphering of Egyptian hieroglyphics was validated by comparing the Demotic, hieroglyphic and ancient Greek texts on the Rosetta stone found in 1799, for example. In 2013, the painting *Sunset at Montmajour* was confirmed as a genuine work by Vincent van Gogh after consulting letters by the artist describing it, and after analysing its chemical composition, colours and themes.

Replicability testing is particularly important for results in humanities disciplines that use empirical methods, and for cornerstone studies. Existing data sets can be reanalysed, or new data can be collected using the same or a modified study protocol (direct or conceptual replication, respectively). Conceptual replication is useful because it allows researchers to triangulate results.

Such testing will depend on preregistration of studies and on providing public access to detailed methods, data-analysis plans and data sets. It is also important to develop and use reporting guidelines for study protocols, publications and data sets. Funding agencies and scientific journals can help by demanding transparency and by funding and publishing replication studies.

Rik Peels, Lex Bouter Vrije Universiteit Amsterdam, the Netherlands. h.d.peels@vu.nl

Use persistent identifiers more

Increasingly, ORCID, DOIs and other identifier systems that are open and community-governed are embedded in scholarly works and information systems, such as papers and citation indices. They could benefit research in many more ways than their current use in unambiguously tracking authors and published output.

Take, for example, the manuscript-submission process. Authors must create a journal account, review submission requirements and upload their manuscript, which probably



Red-footed boobies (Sula sula) nesting in the Chesterfield archipelago.

contains links to other important information. Journals need to find unconflicted reviewers. Payment contacts for open access might be required. By using persistent identifiers, most of the manual processes in this workflow can be semi-automated (see go.nature.com/2lnqibu). Expertise should not be wasted on mundane administrative tasks.

Identifiers can also act as signposts and coordinates, guiding us to information sources and showing connections between research and researchers. They can increase the visibility of a study, its origins and its impact, and indicate where it is hosted and who to ask for access. Contributors, peer reviewers and supporting materials can all be linked to the published article.

An important feature of identifiers is that they afford a wider understanding of the research landscape that does not compromise privacy or 'ownership' of the research itself — pertinent, for example, when the work is ongoing, personal or competitive.

Alice Meadows* ORCID, Brookline, Massachusetts, USA. a.meadows@orcid.org *On behalf of 6 co-signatories; see go.nature.com/2jmznwc for a full list and for competing financial interests.

Keep cruises off remote coral reefs

Just 3% of the world's coral reefs remain in near-pristine condition; about one-third of these are located in the Coral Sea in the South Pacific Ocean. The Chesterfield reef ensemble,

one of the world's largest atolls, is an example. It is part of France's overseas territory of New Caledonia, and its remoteness has so far preserved its wealth of biodiversity. We therefore call for the territory's government to drop its plans to open these precious reefs to the destructive effects of cruise ships and mass ecotourism.

The Chesterfield reefs were spared the 2016 mass-bleaching phenomenon that affected coral reefs around the world. They host the largest seabird colonies in the tropical western Pacific. Indeed, nitrogen from seabird guano may contribute to the resilience of reef-building corals (A. Lorrain et al. Sci. Rep. 7, 3721; 2017).

Comprising a remarkable variety of corals, the reefs host an abundance of diverse fish shoals and species such as the threatened fairy tern (*Sternula nereis*), several endemic marine gastropods and an endemic sea snake (*Hydrophis laboutei*). They are also a nesting site of regional importance for the green sea turtle (*Chelonia mydas*).

Cruise ships will inevitably disrupt the reef and lagoon habitats and fauna. Their hundreds of passengers will lethally disturb breeding seabird colonies, by repeatedly scaring away nesting adults. This could particularly affect the brown booby (Sula leucogaster), the lesser and greater frigatebirds (Fregata ariel and F. minor) and the sooty tern (Onychoprion fuscatus).

Philippe Borsa, Bertrand Richer de Forges French Research Institute for Development, Nouméa, New Caledonia. Julien Baudat-Franceschi Corte, France. philippe.borsa@ird.fr

NEWS & VIEWS

DNA DAMAGE

Stem cells hide from the sun

Adult stem cells reside in niches that maintain, regulate and protect them. Fresh light has now been shed on how the need for protection has driven changes in the locations of these niches during evolution. SEE LETTER P.445

ISABEL BEERMAN

Issue-specific stem cells have the crucial role of maintaining and replenishing all the specialized cells that make up a given tissue or system, so it is essential that their functions are preserved throughout an organism's life. These adult stem cells therefore reside in dedicated microenvironments called stem-cell niches, which help to regulate and protect the stem cells. On page 445, Kapp et al.¹ investigate factors that affect the locations of niches for haematopoietic stem and progenitor cells (HSPCs), which give rise to all blood-cell lineages. Their findings provide insights into the evolutionary drivers of niche location.

Stem-cell niches have several commonalities, regardless of the type of tissue stem cell they sustain. They provide structural support, enable access to molecular signals from either local or remote sources, produce proteins that help tether stem cells to the location and help to regulate stem-cell metabolism². Cues from the niche also restrict unnecessary stem-cell divisions, presumably both to prevent stem-cell exhaustion (in which stem cells lose the ability to regenerate cell lineages) and to minimize the number of genetic mutations that arise during DNA replication, maintaining the fidelity of the genome.

Most information that researchers have about the HSPC niche has been gained from mouse studies. These analyses have revealed that, during development, there are several different sites at which blood lineages are generated³: in mice, HSPCs and their niches move, depending on the developmental stage. Moreover, studies of adult HSPC populations in flies^{4,5}, and in zebrafish^{6,7} and other vertebrates³, have revealed different niche locations in different species — the bone marrow in birds and mammals, the kidney in fish, and the liver in frogs⁸, for example.

A theory published almost 40 years ago states that HSPC niches in terrestrial animals evolved at sites that minimize damage from ionizing radiation⁹, with the assumption that water would have provided aquatic organisms with protection from damaging ultraviolet light. Kapp *et al.* set out to examine whether the need to minimize damage might influence HSPC-niche location in the model organism zebrafish, which tends to inhabit clear bodies

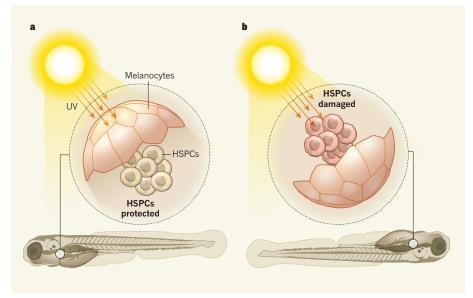


Figure 1 | **Stem cells under an umbrella.** Haematopoietic stem and progenitor cells (HSPCs), which give rise to blood-cell lineages, reside in specialized microenvironments called niches that help to protect them from DNA damage. **a**, Kapp *et al.* have described a population of pigmented cells called melanocytes close to the HSPC niche in zebrafish. The cells act as an opaque umbrella, protecting HSPCs from ultraviolet irradiation. **b**, The authors anaesthetized the fish, causing them to flip on to their backs. The melanocytes no longer protected the HSPCs from UV damage, demonstrating that the pigmented cells provide a physical shield, rather than acting through signalling mechanisms.

of water that offer little UV protection. During embryonic stages, zebrafish are largely transparent, making them well suited to live imaging using fluorescently labelled cells. However, the authors found that visualization of the HSPC population was consistently obscured by pigmented cells called melanocytes.

The researchers used genetic engineering to generate zebrafish lacking melanocytes, and found that HSPCs developed normally, indicating that the pigmented cells are not essential for HSPC maintenance. However, this finding raised the question of what (if any) role the melanocyte population has in relation to HSPCs. Kapp and colleagues performed a series of clever experiments to investigate. They genetically engineered fish to lack melanocyte pigmentation, and showed that UV radiation caused higher levels of DNA damage in these fish than in controls, as well as an increase in HSPC death. Next, they anaesthetized pigmented fish, causing them to flip onto their backs. This removed the protection provided by the physical location of the

melanocytes — again, the HSPCs were rapidly damaged by UV. Thus, melanocytes protect HSPCs by generating a physical opaque shield against UV irradiation (Fig. 1).

Kapp *et al.* next showed that this melanocyte 'umbrella' is present in other species of fish, as well as in frogs, in which melanocytes shield the HSPC niche in tadpoles before the cells migrate to the bone marrow during development. Fish without melanocytes in this location might have had a selective disadvantage during evolution, because their HSPCs would not have been protected by the umbrella and would have been exposed to UV damage. This would have led to decreased numbers of HSPCs, and daily exposure to UV would probably have been lethal.

These findings raise the question of whether aquatic organisms have evolved so that other crucial, vulnerable cells are also physically shielded — either by melanocytes or by other carefully positioned cells — to prevent UV-induced damage. It will be interesting to determine how many other tissue-specific

stem cells have appropriated this type of physical protection in fish.

It remains to be seen whether HSPCs home to locations protected by melanocytes, or whether melanocytes can be recruited or stimulated to proliferate by signals from blood stem-cell niches. Another question is why HSPCs seek out a different niche in terrestrial vertebrates. HSPCs regularly leave their niches and circulate in the blood, whereas other adult stem cells, although capable of movement, tend to be more static. Indeed, Kapp et al. highlighted this mobility in adult Rana frogs, in which the location of HSPCs switches depending on the season — perhaps driven by changing exposure to UV. This mobility could give HSPCs an increased capacity to seek out alternative sites that can better protect them from damage. Bone would completely encapsulate the stem cells, providing a shield from all angles. There might also be other benefits to a niche in the bone marrow, such as lower oxygen levels, which could provide protection from other forms of DNA damage.

Fish have evolved the ability to mitigate UV-induced DNA damage through light-dependent DNA repair¹⁰. The fact that, despite this ability, HSPCs seem to be under evolutionary pressure to seek additional protection highlights the importance of maintaining the fidelity of tissue stem cells. Physical protection might be of particular value in the haematopoietic system, because blood cells have a high turnover rate. If HSPCs are unable to re-establish bloodcell populations, the organism would be likely to die from anaemia or infection.

Finally, it might be predicted that other tissue-specific stem cells that have less migratory potential than HSPCs would be more prone to damage, because they would be less able to move to protective niches. By gaining insight into the evolutionary steps taken to protect haematopoietic stem cells, we might be able to develop strategies to protect these cell types, refining their current niches to better maintain stem-cell potential in longer-lived organisms, including humans.

Isabel Beerman is at the National Institute on Aging, Biomedical Research Center, Baltimore, Maryland 21224, USA.

e-mail: isabel.beerman@nih.gov

- 1. Kapp, F. G. et al. Nature 558, 445-448 (2018).
- Ferraro, F., Celso, C. L. & Scadden, D. Adv. Exp. Med. Biol. 695, 155–168 (2010).
- 3. Orkin, S. H. & Zon, L. I. Cell 132, 631-644 (2008).
- 4. Williams, M. J. J. Immunol. 178, 4711–4716 (2007).
- Evans, C. et al. Adv. Dev. Biol. 18, 259–299 (2007).
 de Jong, J. L. & Zon, L. I. Annu. Rev. Genet. 39,
- 481–501 (2005).

 7. Murayama, E. et al. Immunity **25**, 963–975 (2006).
- Martinez-Agosto, J. A., Mikkola, H. K., Hartenstein, V. & Banerjee, U. Genes Dev. 21, 3044–3060 (2007).
- 9. Horton, J. D. (ed.) *Proc. Symp. Development and Differentiation of Vertebrate Lymphocytes* 1979 (*Dev. Immunol.* Vol. 8; Elsevier, 1980).
- Sinha, R. P. & Häder, D. P. Photochem. Photobiol. Sci. 1, 225–236 (2002).

This article was published online on 13 June 2018.

ASTRONOMY

Missing matter found in the cosmic web

The location of nearly half of the ordinary matter in the Universe is unknown. X-ray observations suggest that this elusive 'baryonic' matter is hidden in the filamentary structure of the cosmic web. SEE LETTER P.406

TAOTAO FANG

e live in a dark Universe: just 5% of it consists of ordinary matter such as that found in atoms, whereas the rest is 'dark' matter and energy that cannot currently be detected directly. However, observations of the nearby Universe suggest that up to 40% of this ordinary matter — which is made up primarily of particles known as baryons — is missing²⁻⁵. Baryonic matter is thought to be distributed through the Universe like a cosmic web, and the missing baryons are predicted to be located in the filamentary structures that connect the web, and in intergalactic space⁴. On page 406, Nicastro et al.⁶ report the detection of the X-ray absorption signatures of baryons in the spectra of a bright background object. The findings might finally reveal a major reservoir for baryonic matter.

Why have the missing baryons been so difficult to detect? One reason is that the density of the baryonic matter in the filaments is extremely low. The other reason is that the high temperature in the filaments causes the most abundant element (hydrogen) to be almost completely ionized — which means that it has no electrons to produce spectral features that could be used to detect it. However, there might be trace amounts of heavier elements such as oxygen, in which a few electrons are bound. These ions can produce detectable (but extremely weak) spectral features,

typically in the X-ray and/or ultraviolet regions of the electromagnetic spectrum.

Nicastro *et al.* observed the X-rays emitted by a special type of astronomical object known as a BL Lacertae (BL Lac) object. These are typically extremely bright, and have no (or very few) intrinsic spectral features — which makes it easy to detect any absorption of their emissions by other objects between them and Earth, such as filaments in the cosmic web.

The BL Lac object studied by the authors is called 1ES 1553+113, and is more than 2,200 megaparsecs away. Nicastro and colleagues observed this target with the European Space Agency's XMM-Newton X-ray Space Telescope over several periods, for a total observation time of about 1.75 million seconds (about 20 days). They thus obtained a spectrum with an extremely high signal-to-noise ratio, which allowed them to perform high-resolution spectroscopy of very weak spectral features (Fig. 1).

The authors discovered two highly statistically significant systems of absorption lines produced by helium-like oxygen (oxygen ions that have only two bound electrons) at redshifts of 0.43 and 0.36. Redshift measures the change in wavelength that occurs when light travels over astronomical distances, and is approximately proportional to the distance of the light-emitting object from Earth. The researchers also performed an optical survey of galaxies along the sight line towards

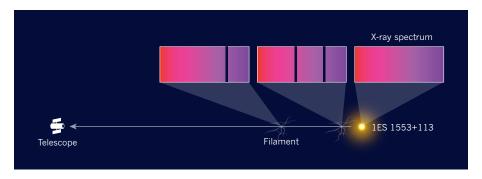


Figure 1 | **The search for baryonic matter.** Nicastro *et al.*⁶ used the XMM-Newton X-ray Space Telescope to detect the emission spectrum of a bright astronomical object called 1ES 1553+113. They observed lines superimposed on the spectrum, which they attribute to X-ray absorption by helium-like oxygen (oxygen ions that have just two bound electrons, not shown) in two filaments of the cosmic web located between the telescope and the emitter. The cosmic web is a massive structure composed of ordinary (baryonic) matter, such as that found in all atoms. If the authors' attribution is correct, then the finding reveals the location of a major reservoir of baryonic matter. Distances and sizes of objects not shown to scale.

1ES 1553+113, and observed a high density of galaxies at the two redshifts associated with the absorption signals. Such densities are characteristic of the filamentary structures of the cosmic web. By combining the X-ray data with measurements of the ultraviolet emissions from 1ES 1553+113, Nicastro et al. estimated the density of the baryons associated with the X-ray absorbing features, and found that they account for 9-40% of the cosmic baryon density — suggesting that these features are a substantial reservoir of the missing baryons.

Weak X-ray absorption lines produced by baryons have been reported a few times before⁷⁻⁸, but most of the results were marginal, and in some cases debatable. What is remarkable about the current work is that it represents the first time both of the expected absorption lines for helium-like oxygen have been detected together (for the absorption system at redshift 0.43, although the statistical significance of one of the lines is marginal). The observation of two absorption lines from the same ion species is typically a good indication that the target ion species has been

One concern is whether the observed X-rayabsorbing systems are truly located between Earth and 1ES 1553+113. The exact redshift of the BL Lac object is unknown; the best available estimate⁶ suggests that it is at least 0.41. This value is less than the redshift of one of the X-ray absorbers, which implies that this absorber is either part of the BL Lac object, or a misidentification of something else. Nicastro et al. argue that both scenarios are unlikely, but an accurate measurement of the redshift of 1ES 1553+113 is needed to resolve this issue.

It is also possible that the X-ray-absorbing systems are in galaxies, rather than in filamentary structures of the cosmic web — similar absorption systems have previously been detected in the Milky Way⁴. Nicastro and coworkers argue that this explanation is unlikely, partly because they did not find large galaxies similar to the Milky Way at the redshifts associated with the absorptions, but also because they did not detect any additional absorption lines from cold ions, which are typically found in galactic disks. These arguments are reasonable, but better observations are needed to rule out this scenario.

This type of observation, requiring more than a million seconds of exposure time, truly pushes the limits of the available instruments. Proposed space missions such as the Hot Universe Baryon Surveyor (go.nature. com/2luj4fa) and the Advanced Telescope for High-Energy Astrophysics (http://sci.esa.int/ athena/) will have much more sensitive X-ray spectrometers, and might eventually provide a complete map of the missing baryons in the cosmic web.

An alternative approach for detecting the missing baryons is to use a phenomenon known as the Sunyaev-Zel'dovich effect,

in which high-energy electrons scatter off photons in the cosmic microwave background (CMB; electromagnetic radiation left over as a remnant of the Big Bang), thereby slightly distorting the CMB spectrum. High-energy electrons outside galaxies, and probably also in the filaments of the cosmic web, could produce such a distortion9, yielding a signal that indicates the presence of baryons. In the meantime, Nicastro and colleagues' findings offer a tantalizing glimpse of where the elusive missing baryons have been hiding.

Taotao Fang is in the Department of Astronomy and the Jiujiang Research Institute, Xiamen University, Xiamen, Fujian 361005,

e-mail: fangt@xmu.edu.cn

- 1. Planck Collaboration. Astron. Astrophys. 594, 13
- 2. Fukugita, M., Hogan, C. J. & Peebles, P. J. E.
- Astrophys. J. **503**, 518 (1998).

 3. Cen, R. & Ostriker, J. P. Astrophys. J. **514**, 1 (1999).
- Bregman, J. N. Annu. Rev. Astron. Astrophys. 45, 221-259 (2007).
- Shull, J. M., Smith, B. D. & Danforth, C. W. *Astrophys. J.* **759**, 23 (2012).
- Nicastro, F. et al. Nature **558**, 406–409 (2018). Fang, T., Marshall, H. L., Lee, J. C., Davis, D. S. &
- Canizares, C. R. *Astrophys. J.* **572**, L127 (2002). Nicastro, F. et al. *Nature* **433**, 495–498 (2005).
- Hernández-Monteagudo, C. et al. Phys. Rev. Lett. 115, 191301 (2015).

CLINICAL ONCOLOGY

Windows open for cancer immunotherapy

Activating immune cells to destroy tumours is an effective strategy for treating an advanced lung cancer — but only for some people. Evidence that this approach has potential in early disease and as a combination therapy has now emerged.

LIZZA E. HENDRIKS & BENJAMIN BESSE

Themotherapy became the standard treatment for lung cancer in the twenti-✓ eth century¹. But in the past 15 years, there has been a drive to improve outcomes for people with this still-deadly disease, either through therapies that target enzymes encoded by genes harbouring cancer-driving mutations, or through immunotherapies, which activate the body's immune system to target tumours. Writing in The New England Journal of Medicine, two groups^{2,3} provide evidence that supports the use of immunotherapies to treat non-small-cell lung cancer (NSCLC) at different stages of the disease.

Tumour cells evade destruction by activating signals known as immune checkpoints, which deactivate immune cells called T cells⁴. Two immune checkpoints are the proteins cytotoxic T lymphocyte antigen 4 (CTLA-4) and programmed cell death 1 (PD-1), which are expressed by T cells themselves. Another, programmed cell death ligand 1 (PD-L1), is produced by tumour cells (Fig. 1).

Antibodies that interact with these proteins to prevent their normal activity, and so reawaken the immune system, are now used to treat metastatic NSCLC — the stage at which the cancer has spread. Antibodies that bind PD-1 or PD-L1 are sometimes successful in patients who have had treatments such as chemotherapy, but whose cancer has nonetheless progressed⁵. Alternatively, the anti-PD-1 antibody pembrolizumab can be used as a first-line treatment for metastatic

NSCLC when the percentage of tumour cells that express PD-L1 is high — these patients respond better to immunotherapy than to chemotherapy⁶.

If such immune-checkpoint-targeted antibodies (ICTs) can improve outcomes for metastatic NSCLC, could they also help to tackle early-stage disease? In the first of the current papers, Forde et al.² carried out a pilot study to investigate whether the anti-PD-1 ICT nivolumab could be used to shrink tumours before surgery, which is a standard treatment for most cases of early-stage NSCLC.

The authors treated 21 patients with 2 doses of nivolumab 2 weeks apart, starting 4 weeks before the planned surgery date. They showed that surgery did not need to be delayed (for example, because of an adverse event with nivolumab) for any patient. The researchers anticipated that four weeks would not be enough time for the reactivated immune system to significantly shrink the tumour. Indeed, imaging revealed significant shrinkage in tumours in only two patients before surgery. However, examination of the surgically removed tumours revealed that 45% had undergone a major response to the ICT — less than 10% of the tumour cells remained alive. ICTs, unlike chemotherapy, cause inflammation and scar-tissue formation in tumours, and can therefore sometimes cause tumour growth. However, the researchers found that even two tumours that showed such growth had undergone a strong pathological response.

This level of efficacy is impressive, but needs to be further investigated by

Figure 1 | Reawakening the immune system through different mechanisms. a, Tumour cells express the protein PD-L1 on their surfaces. PD-L1 binds to PD-1 receptors on the surface of immune cells called T cells, inactivating the cells and so preventing them from targeting tumour cells for destruction. The antibody nivolumab prevents this interaction, and so reactivates T cells to destroy tumour cells. **b**, T cells can also be activated through interactions

with another type of immune cell, dendritic cells, between the proteins CD28 and B7, and TCR and MHC. The T-cell protein CTLA4 binds to B7 to block its interaction with CD28, preventing T-cell activation. The antibody ipilimumab blocks CTLA4, reactivating T cells to destroy tumour cells. Two studies^{2,3} now provide evidence that nivolumab and ipilimumab can be used to help treat non-small-cell lung cancer, at either early or advanced stages.

in-depth examination of the tumour specimens collected. Moreover, the optimal duration of nivolumab treatment remains to be determined, because delayed responses to ICTs can occur⁷. Of note, one of the patients enrolled experienced severe acute toxicity to nivolumab. Possible long-term side effects also need to be considered, because Forde *et al.* followed their patients for only a median of 12 months after surgery. Phase III trials are therefore now essential.

Forde and colleagues found that the number of mutations in each tumour's genome correlated with whether that tumour had a major pathological response to the ICT. And another study⁸ has found that a high number of mutations in a tumour — the tumour mutational burden (TMB) — is associated with an improved outcome in patients treated with a combination of ICTs. So, can TMB be used to predict which patients with advanced NSCLC would benefit from immunotherapy? In the second of the current papers, Hellmann et al.³ investigated this possibility, as part of a phase III trial called CheckMate 227.

The authors selected patients who had untreated, advanced-stage NSCLC. They assigned patients to one of four treatment groups — nivolumab; chemotherapy plus nivolumab; chemotherapy; or nivolumab plus the anti-CTLA4 antibody ipilimumab. For the last two groups, they analysed how a high TMB of at least 10 mutations per megabase of DNA affected progression-free survival — the time before the tumour begins to grow or spread once more.

Hellmann and colleagues found that median progression-free survival for patients who had a high TMB was 7.2 months for nivolumab plus ipilimumab, compared with 5.5 months for chemotherapy. After one year, there was no progression in 42.6% of patients who received the combination immunotherapy, compared with 13.2% of those who received chemotherapy alone. However, people with a low TMB did

not benefit from combination ICT. These results are remarkable — the first positive results for a combination ICT predicted using TMB.

What will the impact of these findings be, in an already crowded treatment field? One issue is the feasibility of using TMB to find suitable patients. Out of the 1,739 patients enrolled by Hellmann *et al.*, only 1,004 could be evaluated for TMB, mainly because the amount of tissue available or the quality of the DNA extracted was inadequate. The analysis of progression-free survival in patients with a high TMB involved only 299 people. Furthermore, there are multiple tests for TMB that analyse different

"The authors found that the number of mutations in each tumour's genome correlated with whether that tumour had a major response to the immunotherapy."

genomic regions, and multiple cut-offs for high TMB classification — no cross-test validation exists. Other drawbacks of TMB analysis are the cost; the amount of tissue needed; and the fact that the analysis takes about two weeks. Finally, some of the tumours that had high TMB

in Hellmann and colleagues' study also had high levels of PD-L1 expression. In these cases, single-agent immunotherapy is effective, and is less toxic than the combination treatment⁶.

It should be noted that two recently reported phase III trials^{9,10} found that the combination of chemotherapy and an ICT is superior to chemotherapy, whatever the level of PD-L1 expression (and possibly whatever the TMB). It is expected that these studies will establish a new standard of care. It will be hard for ICT combinations to compete.

ICTs can also have a dark side. One study¹¹ found that 14% of people with NSCLC who were treated with ICTs developed hyperprogressive disease — an increase in tumour growth rate compared to the rate when

patients were given a previous treatment. Hyperprogression is associated with poor survival rates. In the first few months of Hellmann and colleagues' trial, progression-free survival was higher for the chemotherapy arm than for the combination ICT, possibly suggestive of hyperprogression in some patients on the combination ICT. Whether such a pattern might arise in early-stage NSCLC, such as that examined by Forde *et al.*, should also be carefully monitored. Finally, ICTs could harm our health-care systems through their cost — more than US\$10,000 per month for a combination treatment¹². We will need to monitor whether such a cost is sustainable.

Lizza E. Hendriks and Benjamin Besse are in the Department of Cancer Medicine, Institut d'Oncologie Thoracique, Gustave Roussy, 94805 Villejuif, France. L.E.H. is also in the Department of Pulmonary Diseases, Maastricht University Medical Center, the Netherlands. B.B. is also at the Université Paris Sud, Université Paris-Saclay. e-mails: lizza.hendriks@mumc.nl; benjamin.besse@gustaveroussy.fr

- NSCLC Meta-Analyses Collaborative Group. J. Clin. Oncol. 26, 4617–4625 (2008).
- Forde, P. M. et al. N. Engl. J. Med. 378, 1976–1986 (2018).
- 3. Hellmann, M. D. et al. N. Engl. J. Med. **378**, 2093–2104 (2018).
- 4. Hoos, A. Nature Rev. Drug Discov. 15, 235–247 (2016).
- 5. Bianco, A., Malapelle, U., Rocco, D., Perrotta, F. & Mazzarella, G. Curr. Opin. Pharmacol. **40**, 46–50 (2018).
- Reck, M. et al. N. Engl. J. Med. 375, 1823–1833 (2016).
- 7. Brahmer, J. et al. N. Engl. J. Med. **373**, 123–135 (2015).
- Hellmann, M. D. et al. Cancer Cell https://doi. org/10.1016/j.ccell.2018.03.018 (2018).
- Gandhi, L. et al. N. Engl. J. Med. https://doi. org/10.1056/NEJMoa1801005 (2018).
 Reck, M. et al. Ann. Oncol. 28 (Suppl. 11), xi31 (2017).
- 10. Reck, M. et al. Ann. Oncol. **28** (Suppl. 11), xl31 (2017) 11. Ferrara, R. et al. J. Thorac. Oncol. **12** (Suppl. 2), S1843–S1844 (2017).
- 12.NICE. https://www.nice.org.uk/guidance/ta400 (2016).

This article was published online on 13 June 2018.

MARINE GEOLOGY

Sea-level rise could overwhelm coral reefs

An assessment of the capacity of coral reefs to grow fast enough to keep up with projected rises in sea level finds that most reefs will fall behind if nothing is done to restore them. SEE ARTICLE P.396

ILSA B. KUFFNER

oral reefs are famous for housing biodiversity and attracting tourists, and the economic benefits that reefs provide for tropical, coastal communities around the globe measure in the billions of dollars¹. One of the main services provided by reefs is that they act as natural breakwaters (Fig. 1), protecting shorelines and human-built infrastructure from storms. On page 396, Perry et al.² report a detailed analysis of the ability of coral reefs in two ocean basins to keep growing upwards in the face of the ecological degradation they have already experienced, and taking into account future sea-level rise. The findings show that, as living coral populations wane, their capacity to build reefs might be diminished to the point at which the reef community fails to keep up with the rising ocean surface.

Corals are simple invertebrates that are related to sea anemones and jellyfish, but a trait that sets them apart is their ability to create rock from sea water. Each year, corals add a new layer of calcium carbonate on top of their existing exoskeletons, growing larger and intertwining over thousands of years to form a coastal barrier capable of quelling

enormous amounts of wave energy. Corals are under threat from warming oceans and from an onslaught of localized environmental pressures, which collectively cause coral bleaching, slower growth, disease and death³. If there are not enough corals alive to keep a reef growing, then erosion takes over and the reef loses elevation.

There are some uncertainties associated with understanding the fate of coral reefs as geological structures. Coral growth does not translate millimetre for millimetre into vertical reef growth. Many constructional and erosional processes are at work simultaneously, adding to and subtracting from the net amount of calcium carbonate produced or lost (the carbonate budget), and determining whether a reef builds or winnows away. In previously published work, Perry and colleagues⁴ were among the first to use field data that account for the organisms responsible for reef building (corals and calcifying algae) and reef breakdown (excavating parrotfish, sea urchins and reef-infesting sponges) in budgeting projections of reef growth and destruction.

In the current work, Perry *et al.* take those budgeting efforts a step further by combining them with projections of sea-level rise under

two scenarios published in the Fifth Assessment Report⁵ from the Intergovernmental Panel on Climate Change (IPCC). What they find is not encouraging: 16 reef areas in the tropical western Atlantic Ocean and 6 in the Indian Ocean are barely keeping up with the present sea level. Even worse, only 9% of the 202 reefs they assessed have the capacity to keep up with the rates of sea-level rise associated with even the more optimistic of the two scenarios (IPCC Representative Concentration Pathway 4.5), which predicts that atmospheric greenhousegas emissions will peak around 2040.

The authors acknowledge that their accounting does not adequately address some reef processes that are important in the budget. If you used a rotary drill to take a peek inside a reef, you would find that only about half of the structure is composed of intact coral skeletons; the rest is either void space or reef detritus, including rubble and sediment⁶. The processes that control the breakdown of reefs and determine whether the resulting material fills the cracks and crevices or gets swept away are not well studied. Although Perry and colleagues' budget did account for the biologically mediated erosion responsible for producing reef rubble and sediments, they did not factor in the chemical dissolution of carbonates, nor evidence suggesting^{7,8} that both of these processes will be accelerated by increased levels of carbon dioxide absorption by the ocean (ocean acidification).

Additionally, the transport of loose sediment away from reefs and into deeper water is largely driven by sporadic storms. This makes it difficult to estimate an average rate at which sediment contributes to reef building. Moreover, sediment transport away from reefs might increase as cyclones become more intense as a result of ocean warming⁹ — a factor that was also not considered by the authors. Taken together, the processes not accounted for by Perry and colleagues could mean that the projections of reef-building rates are, if anything, too optimistic.

The implications of the study are dire: many island nations and territories are set to quickly lose crucial natural resources responsible for coastal defence. Prompt action is warranted to slow and reverse this loss. Fortunately, reef restoration has come a long way since the twentieth century, when piles of discarded car tyres and engine blocks were used as artificial reefs. Restoration using live corals farmed in offshore nurseries is fast becoming more common and feasible as coral-gardening techniques have been streamlined. So far, these efforts have been driven largely by conservation organizations and hoteliers, but reef-restoration programmes are poised to benefit from initiatives that coordinate restoration practitioners, scientists, governments, resource managers and local communities (see, for example, go.nature. com/2rljaqh).

The feasibility and efficacy of enhancing coastal and community resilience through



Figure 1 Natural breakwaters. Coral reefs, such as this one in the Red Sea off the coast of Saudi Arabia, protect shorelines from stormy seas. Perry *et al.*² report that most coral reefs might not be able to grow fast enough to keep up with future sea-level rises, putting crucial coastal defences at risk.

live-coral planting have not been quantified, but Perry et al. provide convincing evidence that the time has come to make reef restoration a priority. A recent analysis 10 indicates that ecological restoration projects aimed at protecting shorelines might be more costeffective than conventional projects that use engineered concrete structures. Although it is uncertain how much time we can buy for coral reefs through restoration, such projects might extend the existence of the reefs long enough to bridge the gap until global efforts start to decrease the concentration of atmospheric

greenhouse gases, thereby slowing the rates of global warming and sea-level rise. ■

Ilsa B. Kuffner is at the US Geological Survey, St. Petersburg Coastal and Marine Science Center, St. Petersburg, Florida 33701, USA. e-mail: ikuffner@usgs.gov

- Costanza, R. et al. Nature **387**, 253–260 (1997). Perry, C. T. et al. Nature **558**, 396–400 (2018). Hughes, T. P. et al. Nature **546**, 82–90 (2017).
- Perry, C. T. et al. Coral Reefs 31, 853–868 (2012)
- Church, J. A. et al. in Climate Change 2013: The Physical Science Basis. Contribution of Working

- Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (eds Stocker, T. F. et al.) Ch. 13 (Cambridge Univ. Press, 2013).
- 6. Hubbard, D. K., Miller, A. I. & Scaturo, D.
- J. Sedimentol. Petrol. **60**, 335–360 (1990). Tribollet, A., Godinot, C., Atkinson, A. & Langdon, C. Global Biogeochem. Cycles 23, GB3008 (2009)
- Silbiger, N. J. & Donahue, M. J. Biogeosciences 12,
- Swearer, S. E. Glob. Change Biol. 24, 1827-1842 (2018)

This article was published online on 13 June 2018.

NANOSCIENCE

Frictionless when flat

Gas transport in nanoscale channels that have perfectly flat walls has been found to be frictionless, challenging the classical theory of gas flow. The findings might enable new devices for gas separation and flow control. See Letter P.420

CHUANHUA DUAN

'n classical physics, frictionless gas flow through channels is a phenomenon that can occur only under ideal conditions. A key requirement is that the gas molecules undergo specular reflection from the channel walls they rebound so that their angle of incidence is the same as the angle of reflection. In reality, gas molecules are thought to rebound from walls in all directions, a behaviour known as diffuse reflection. But can diffuse reflection be switched to specular reflection to achieve frictionless gas transport, and, if so, what are the fundamental requirements for this? On page 420, Keerthi et al.1 answer these questions by investigating gas transport through two-dimensional, nanometre-scale channels that have atomically flat walls.

Achieving specular reflection for gas molecules seems an impossible task. Even the best, artificially polished surface is randomly bumpy at the tiny scales associated with gas molecules (which are about 10^{-9} – 10^{-10} metres in diameter). It is this bumpiness that is thought to make gas molecules rebound in all directions, causing diffuse reflection. The concept of diffuse reflection underpins Knudsen theory, which has generally provided a good description of gas transport through channels. But what happens if the channel surfaces are genuinely flat?

To answer this question, Keerthi et al. prepared nanochannels^{2,3} from materials that consist of stacked, 2D layers of atoms: graphite, hexagonal boron nitride (h-BN) and molybdenum disulfide (MoS₂). These materials can be cleaved to form 2D crystals that are as thin as a single layer of atoms, and have atomically flat surfaces. 2D crystals made from different materials can easily stack together, because

strong van der Waals forces form between their flat surfaces⁴. The authors made their nanochannels by sandwiching a 2D crystal containing a narrow gap between two other 2D crystals (Fig. 1). The channels are well sealed, and their heights can be precisely controlled by the number of layers of atoms in the middle 2D crystal.

The researchers measured the flow of gas through the nanochannels under low pressure. Under these conditions, the gas molecules collide mostly with the surfaces of the channels, rather than with each other. The authors found that the permeability of helium through nanochannels made from graphite or h-BN, whose surfaces are flat even at scales of 1 ångström (10⁻¹⁰ m), is ten to several hundred times higher than predicted by Knudsen theory. Furthermore, they observed that the highest helium permeability measured in graphite nanochannels was independent of the channel length, indicating that there is no momentum loss in the channels and that frictionless flow has been achieved.

By contrast, the permeability of helium through MoS₂ nanochannels was the same as that predicted by Knudsen theory. This is

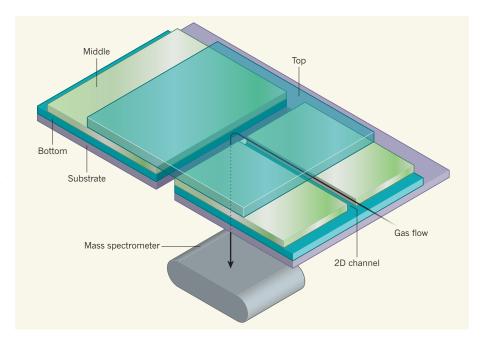


Figure 1 | Measuring gas flow though 2D channels. Keerthi et al. 1 prepared nanochannels from materials that can be cleaved to produce 2D crystals just one or a few layers of atoms thick. The nanochannels were made by sandwiching a 2D crystal (middle layer) containing a narrow gap between two other 2D crystals (top and bottom layers). The authors then measured the flow of gas — helium, hydrogen or deuterium — through the channels using a mass spectrometer. In channels made of graphite and hexagonal boron nitride, they observed much higher gas flow than is predicted by classical theory; this flow was confirmed to be frictionless in the graphite channels. However, gas flow conforming to classical theory was observed in channels made of molybdenum disulfide. The properties of the flow seem to depend on the roughness of the channel surfaces as 'perceived' by the gas molecules.

because MoS_2 is rougher at atomic scales than is graphite (and h-BN). Through computational modelling, the authors found that MoS_2 surfaces have bumps around 1 Å in height, which is comparable to the diameter and the de Broglie wavelength of helium molecules (all matter can exhibit wave-like behaviour, and the de Broglie wavelength is the wavelength associated with that behaviour). In other words, the MoS_2 nanochannels are too bumpy for specular reflection.

Keerthi and colleagues' findings prove that complete specular reflection and frictionless gas transport can occur in nanochannels that have perfectly flat surfaces. This is an exciting discovery, because previous studies^{5,6} have reported only partial specular reflection. Even more intriguingly, the authors make several unexpected observations, some of which cannot be explained by classical physics.

The first and most important of these observations is that frictionless gas transport is affected by the matter waves of the gas molecules. The authors found that the permeability of deuterium (D₂) in graphite nanochannels is much lower than that of hydrogen (H₂), its lighter isotopic counterpart, even though Knudsen theory predicts the opposite. This is because deuterium molecules have a smaller de Broglie wavelength than do hydrogen molecules, and therefore 'see' the channel walls as being rougher, even though the two types of molecule have the same diameter and interact in the same way with the channel walls. The authors also showed that computational simulations of gases that represent classical molecule-wall interactions, but not quantum effects (that is, the effects of matter waves), predict only partial, rather than complete, specular reflection in graphite and h-BN nanochannels. This suggests that quantum effects must contribute to specular reflection.

The other interesting observation is that the permeability of helium in graphite and h-BN channels varies unexpectedly with channel height: it initially increases, then decreases as the channel height increases, reaching a maximum value for heights of four atom layers. This behaviour is at odds with conventional thinking that complete specular reflection is not affected by channel height. Keerthi et al. speculate that the height dependence results from the interplay between two effects: small channels have relatively small 'capture zones' for incident gas molecules at their entrances, whereas hydrocarbons from the surrounding air can be adsorbed to larger channels during channel fabrication, roughening the atomically flat surface. Neither of these effects is included in existing models of gas flow, but they seem to have key roles in determining permeability in the real world.

The new findings call for a re-examination of the classical physics of gas dynamics at low pressure and its correlation with quantum mechanics. However, more experiments with other gases in graphite and h-BN nanochannels are needed to further unravel the influence of

molecular diameter and de Broglie wavelength on specular reflection, given that larger diameters typically correspond to smaller de Broglie wavelengths. Moreover, a quantitative comparison of gas transport through rectangular nanochannels and through circular nanotubes made of the same material is needed to evaluate the effect of channel curvature.

In addition, the factors that cause the degradation of specular reflection should be investigated, such as the affinity of gas molecules for channel walls. The variation of gas permeability as a function of confinement and temperature should also be measured for both specular and diffuse reflection. In parallel with the experimental work, more simulations or theoretical work that consider quantum effects are needed, to quantitatively understand and predict the properties of frictionless gas transport.

Such research potentially offers comprehensive insight into the nature of gas transport through channels and at low pressures. Knowledge of such gas transport has found extensive application in studies of the aerodynamics of space vehicles, in micro-electromechanical systems, and in shale-gas extraction^{7,8}. Further research might also shed light on how laminar

membranes can be made from 2D materials for separating mixtures of gases, thereby improving separation efficiency while reducing energy consumption⁹. Finally, frictionless gas transport through channels that are asymmetrically constricted¹⁰ could enable gas-flow rectification¹¹, a process that might allow the development of new pumps, valves and other devices for controlling gas flow. ■

Chuanhua Duan is in the Department of Mechanical Engineering, Boston University, Boston, Massachusetts 02215, USA. e-mail: duan@bu.edu

- 1. Keerthi, A. et al. Nature **558**, 420–424 (2018).
- 2. Esfandiar, A. et al. Science **358**, 511–513 (2017).
- 3. Radha, B. et al. Nature 538, 222-225 (2016).
- Geim, A. K. & Grigorieva, I. V. Nature 499, 419–425 (2013).
- Holt, J. K. et al. Science 312, 1034–1037 (2006).
- Majumder, M., Chopra, N. & Hinds, B. J. ACS Nano 5, 3867–3877 (2011).
- 7. Wu, L. et al. J. Fluid Mech. **822**, 398–417 (2017).
- 8. Shen, C. Rarefied Gas Dynamics: Fundamentals, Simulations and Micro Flows (Springer, 2005).
- Liu, G., Jin, W. & Xu, N. Angew. Chem. Int. Edn 55, 13384–13397 (2016).
- 10.Zhang, P. & Hung D. *J. Appl. Phys.* **115**, 204908 (2014).
- 11. Groisman A. & Quake, S. R. Phys. Rev. Lett. **92**, 094501 (2004).

STEM CELLS

Intestinal-niche conundrum solved

The cellular microenvironment required to sustain adult intestinal stem cells has long been controversial. Cells that release proteins needed for intestinal-tissue renewal have now been defined. SEE LETTER P.449

LINDA C. SAMUELSON

he human small intestine and colon are maintained throughout life by tissue-resident stem cells, which renew the gut lining at an astounding rate by generating billions of cells every day. The self-renewal, proliferation and differentiation of these intestinal stem cells are prompted by molecular signals from nearby cells called niche cells. Over the past decade, stem-cell biologists have debated the identity of the intestinal niche cells¹. Two papers in *Nature* (one by Shoshkes-Carmel et al.² published earlier this year, and the other by Degirmenci et al.³ on page 449) now identify a niche-cell population that provides a signal essential for stem-cell renewal.

Intestinal cells are arranged in a strict spatial layout. Stem cells are found at the base of pit-like structures in the gut wall, called crypts (Fig. 1). They produce highly proliferative progenitor cells, which differentiate into the various mature epithelial-cell types that make

up the gut lining as they move away from the crypt base.

By contrast, Paneth cells are differentiated epithelial cells that move down from the progenitor zone to the crypt base. The close physical association between Paneth cells and stem cells has led to the proposal that Paneth cells have niche function, promoting stem-cell self-renewal. Providing support for this idea, a study has shown that Paneth cells enhance intestinal stem-cell growth in culture, and that intestinal stem-cell numbers are reduced in mice in which Paneth-cell numbers are artificially depleted⁴. However, other studies have come to the opposite conclusion, showing normal stem-cell function after Paneth-cell loss^{5,6}. Furthermore, Paneth cells are not normally found in colonic crypts. Thus, the status of Paneth cells as niche cells is controversial.

Searching for cells that produce the telltale proteins that promote stem-cell self-renewal is one way of identifying the niche-cell population. The main signalling pathways for

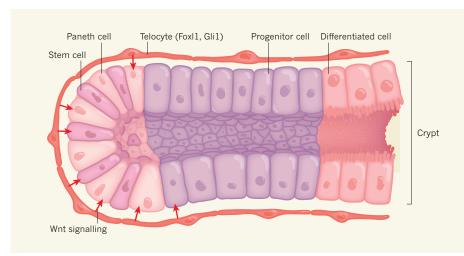


Figure 1 | The intestinal stem-cell niche. Intestinal stem cells generate highly proliferative progenitor cells that differentiate into the various epithelial-cell types that make up the gut lining. The stem cells are located in the base of structures called crypts, and are interspersed with differentiated epithelial cells called Paneth cells. Wnt proteins secreted from a previously unknown cell type trigger Wnt signalling at the crypt base, and these proteins are essential for stem-cell maintenance. Two studies^{2,3} provide evidence that a specialized non-epithelial cell called a telocyte emits Wnts. The telocytes are characterized by production of the proteins Foxl1 and Gli1, and have long extensions that form a sub-epithelial network to support tissue renewal.

stem-cell maintenance are Notch⁷ and Wnt⁸. The Wnt signalling pathway is of particular interest, because elevated signalling in stem cells leads to unchecked proliferation and tumour formation⁸. Indeed, mutations that activate the Wnt signalling pathway are associated with most human colon cancers⁹.

Wnts are secreted proteins that bind to cell-surface receptors to activate the Wnt signalling pathway in neighbouring cells. They are produced by many different intestinal cells, including Paneth cells. However, genetic depletion of Wnts in all intestinal epithelial cells does not alter stem-cell function or crypt structure ¹⁰. Therefore, intestinal epithelial cells are not an essential source of Wnt. This implies that there must be an alternative, nonepithelial-cell source for this niche factor. But what is that source?

An intestinal population of connectivetissue cells called stromal cells, which produce the protein Foxl1, is required for crypt-cell proliferation in adult mice¹¹. Shoshkes-Carmel et al. took advantage of this fact to identify a Wnt source. They used genetic engineering to delete the gene Porcn, which encodes a protein required for Wnt secretion, in Foxl1producing cells in mice. This resulted in rapid, catastrophic crypt collapse, a decrease in the number of intestinal stem cells, and defects in epithelial-cell proliferation. Extensive epithelial degradation was apparent in both the small intestine and colon three days after *Porcn* deletion. Thus, the Foxl1-producing stromal cells are the elusive Wnt-producing niche cells.

Degirmenci *et al.* took a similar approach, but analysed stromal cells that produced a different protein, Gli1. The same research group previously demonstrated that this cell population expressed Wnts¹². The authors deleted the

gene *Wls*, which (like *Porcn*) encodes a protein required for Wnt secretion, in Gli1-producing cells in mice. This led to crypt collapse in the colon. No effects on the small intestine became apparent until the researchers deleted *Wls* in both epithelial cells and Gli1-producing stromal cells. Surprisingly, the response was relatively slow in both cases — stem-cell loss and crypt collapse took two to three weeks.

Wnt depletion has a more limited, slower effect in Gli1-producing cells than in Foxl1producing cells, but the reason for this is unclear. Perhaps Wnt secretion was not completely blocked in the Gli1-producing cells in the small intestine, and the slower effect was due to the long half-life of the Wls protein¹². Regardless of the differences, the fact that blocking Wnts in each cell population leads to stem-cell loss suggests that these stromal cells are the elusive Wnt-producing niche cells. The authors of the two papers have shown that there is physical overlap between the Foxl1-producing and Gli1-producing stromal cells, suggesting that each study identified the same cell population. However, single-cell analyses revealed subpopulations within each of these stromal-cell types. Whether only certain subpopulations have niche function remains to be defined.

Both groups found that the Wnt-expressing niche cells were positioned close to intestinal epithelial cells, in a prime location to affect stem-cell function. Shoshkes-Carmel *et al.* performed high-resolution microscopy, which revealed that these cells are telocytes — thin cells with long protrusions called telopodes, which form a 3D network underlying the epithelial cells throughout the gut. Both groups showed that these telocytes might be signalling hubs, because they express, in addition to Wnts, several other niche factors involved

in stem-cell function and tissue renewal. Moreover, telopodes have the potential to interact with immune cells, blood vessels and nerves¹³. Telocytes might therefore be central coordinators of intestinal renewal beyond their role in Wnt signalling.

Wnt signalling is highest at the base of the crypts, where stem cells reside1. Shoshkes-Carmel et al. showed that telocytes express different levels of Wnts and Wnt inhibitors along the length of the crypt, with higher levels of Wnt protein at the crypt base enabling localized activation of Wnt signalling in stem cells. Whether this compartmentalization reflects reciprocal signalling interactions between stem cells and telocytes is an interesting question — if so, it could indicate that telocytes are responsive to stem-cell status. In support of this idea, Degirmenci et al. showed that the numbers of Gli1-producing cells increased after colon damage, suggesting that this stromal population might adapt to restore homeostasis after damage.

Identification of telocytes as Wnt niche cells has resolved the controversy over which Wnt source regulates intestinal maintenance. A key next step will be to characterize the telocytes' role in intestinal regeneration after injury, or in diseases that affect stem-cell proliferation, such as colon cancer. Further analysis will also be needed to determine whether a single telocyte population is responsible for niche function, or if different populations orchestrate the many stromal growth factors and inhibitors that regulate intestinal stem-cell function. A deeper understanding of how telocytes regulate intestinal stem cells is likely to provide insights into mechanisms of normal intestinal-tissue renewal and regeneration, and dysfunction associated with intestinal disease.

Linda C. Samuelson is in the Department of Molecular and Integrative Physiology, University of Michigan Medical School, Ann Arbor, Michigan 48109–2200, USA. e-mail: lcsam@umich.edu

- Mah, A. T., Yan, K. S. & Kuo, C. J. J. Physiol. (Lond.) 594, 4837–4847 (2016).
- 2. Shoshkes-Carmel, M. et al. Nature **557**, 242–246 (2018)
- Degirmenci, B., Valenta, T., Dimitrieva, S., Hausmann, G. & Basler, K. Nature 558, 449–453 (2018).
- 4. Šato, Ť. et al. Nature **469**, 415–418 (2011).
- Kim, T.-H., Escudero, S. & Shivdasani, R. A. Proc. Natl Acad. Sci. USA 109, 3932–3937 (2012).
- Durand, A. et al. Proc. Natl Acad. Sci. USA 109, 8965–8970 (2012).
- VanDussen, K. L. et al. Development 139, 488–497 (2012).
- 8. Barker, N. et al. Nature 475, 608–611 (2009)
- Yaeger, R. et al. Cancer Cell 33, 125–136 (2018).
 San Roman, A. K., Jayewickreme, C. D., Murtaugh, L. C. & Shivdasani, R. A. Stem Cell Rep. 2,
- 127–134 (2014). 11.Aoki, R. et al. Cell. Mol. Gastroenterol. Hepatol. **2**, 175–188 (2016).
- 12.Valenta, T. et al. Cell Rep. **15**, 911–918 (2016).
- Cretoiu, D., Radu, B. M., Banciu, A., Banciu, D. D. & Cretoiu, S. M. Sem. Cell Dev. Biol. 64, 26–39 (2017).

This article was published online on 6 June 2018.



Antarctic ice shelf disintegration triggered by sea ice loss and ocean swell

Robert A. Massom^{1,2}*, Theodore A. Scambos³, Luke G. Bennetts⁴, Phillip Reid^{2,5}, Vernon A. Squire⁶ & Sharon E. Stammerjohn⁷

Understanding the causes of recent catastrophic ice shelf disintegrations is a crucial step towards improving coupled models of the Antarctic Ice Sheet and predicting its future state and contribution to sea-level rise. An overlooked climate-related causal factor is regional sea ice loss. Here we show that for the disintegration events observed (the collapse of the Larsen A and B and Wilkins ice shelves), the increased seasonal absence of a protective sea ice buffer enabled increased flexure of vulnerable outer ice shelf margins by ocean swells that probably weakened them to the point of calving. This outer-margin calving triggered wider-scale disintegration of ice shelves compromised by multiple factors in preceding years, with key prerequisites being extensive flooding and outer-margin fracturing. Wave-induced flexure is particularly effective in outermost ice shelf regions thinned by bottom crevassing. Our analysis of satellite and ocean-wave data and modelling of combined ice shelf, sea ice and wave properties highlights the need for ice sheet models to account for sea ice and ocean waves.

The recent abrupt, rapid and catastrophic large-scale disintegrations of ice shelves along the Antarctic Peninsula is of major concern because this process indirectly contributes to sea level rise from the Antarctic Ice Sheet. Ice shelves and floating outlet glaciers fringe 74% of the ice sheet coastal margin¹, and play a crucial part in moderating the discharge of grounded ice into the ocean^{2,3}. Until now, disintegration events have been attributed to a combination of (1) enhanced regional warming^{4,5}, leading to increased (wide-scale) surface meltwater ponding and flooding⁶ and crevasse enlargement by hydrofracture⁵, (2) changes in ice material strength due to infiltration of warmth by meltwater percolation⁷, plus thinning and increased basal melt^{8,9}, (3) brine infiltration¹⁰ and (4) glaciological factors involving more localized outer-margin fracturing by bending stresses induced by buoyancy forces^{7,10} and the propagation of existing structural weaknesses such as crevasses and rifts^{9,11–13}. These processes probably all contribute at some level to compromising ice shelf structural integrity and to shelf destabilization.

An additional characteristic of the ice shelf disintegration events is an abrupt transition from a quasi-stable ice shelf front to catastrophic large-scale disintegration. Here, we propose a trigger mechanism involving regional loss of pack ice or shelf-fringing landfast ice, allowing storm-generated ocean swell to flex the outer margins of the shelves and their pre-existing (probably water-filled) fractures, resulting in outer-margin calving and then sudden extensive collapse.

Work dating back to the 1970s^{14–16} shows that ocean waves (ranging from short-period swells to very-long-period transoceanic infragravity or irregularly occurring tsunami waves) impose flexural strains on Antarctic ice shelves, with the potential to induce crevasse and rift propagation and calving^{17–20} (see Methods for a discussion of very-long-period wave and tidal effects). Notably, a study of the Erebus Glacier Tongue in the Ross Sea²¹ implicated the removal of landfast sea ice combined with ocean swell in a relatively small-scale iceberg calving event (in March 1990). Swell is defined as relatively long-period surface-gravity waves that are generated by distant weather systems and are no longer growing or being sustained locally by the wind, as opposed to locally generated wind waves.

Here we present evidence that regional loss of sea ice before and during the disintegration events allows storm-induced long-period (10–20 s) ocean swells to reach exposed ice shelf fronts that have been preconditioned for calving by extensive fracturing and meltwater flooding. These swells excite flexural oscillations in the outer ice shelf margin, imparting sufficient strain to cause cumulative shelf-front fatigue, fracture amplification and ultimately calving. This triggers wider disintegration of the interior shelf areas driven by extensive surface flooding and hydrofracture (Fig. 1). Interior areas are susceptible to hydrofracture (see case examples described below) as a result of very low internal compressive stress arising from thinning and loss of shear stress at the margins and pinning points (see, for example, refs 9-11,12). Moreover, we show that swells are strongly attenuated by the presence of extensive sea ice (see Methods), substantially reducing their destructive effect. Thus, sea ice in the vicinity of weakened or flooded shelves acts as a protective buffer, and its loss is potentially the ultimate cause of rapid ice shelf disintegration in the cases examined.

We develop a conceptual model of the sequence and common factors involved (Fig. 1) by examining well characterized disintegration events on three ice shelves (Fig. 2). These events are the Larsen A Ice Shelf in 1995 (with an areal loss of about 860 km² in just the 2 days between 28 and 30 January 1995, and a total loss of 1,600 km² in 39 days 13); the Larsen B Ice Shelf in 2002 (with an areal loss of 3,320 km² from 31 January to 17 March 2002 (with an areal loss of shelf in February–March and May–July 2008 (with an areal loss of about 800 km²) and in March–April 2009 (with an areal loss of about 1,450 km²) 10,22 . The Larsen B Ice Shelf collapse in 2002 removed an area of ice shelf that had remained intact for the previous 11.5 millennia 23 , and led to a threefold to eightfold acceleration in grounded ice discharge from its tributary glaciers 24,25 .

Prerequisites for disintegration

In the case of the three ice shelves examined, four common factors are essential for disintegration, and form the basis of our conceptual model (Fig. 1). These essential prerequisites are: (I) extensive surface flooding and hydrofracture; (II) reduced sea ice in the region of the ice

¹Australian Antarctic Division, Kingston, Tasmania, Australia. ²Antarctic Climate and Ecosystems CRC, Hobart, Tasmania, Australia. ³National Snow and Ice Data Center, University of Colorado, Boulder, CO, USA. ⁴School of Mathematical Sciences, University of Adelaide, Adelaide, Australia, Australia. ⁵Australian Bureau of Meteorology, Hobart, Tasmania, Australia. ⁶University of Otago, Dunedin, New Zealand. ⁷Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO, USA. *e-mail: rob.massom@aad.gov.au

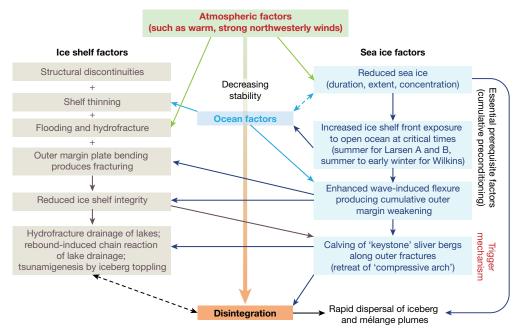


Fig. 1 | Linkages and processes in the structural weakening over time, then abrupt disintegration, of the Larsen A and B and Wilkins ice shelves. These involve change in sea ice and are underpinned by and coupled to atmosphere and ocean processes. In the case of the Wilkins Ice

Shelf, sea ice refers to both pack ice and stationary landfast ice attached to the ice shelf fronts. Linkages and processes marked with dashed lines indicate potential feedbacks, for example, between increased open water (lack of sea ice) and ocean processes.

shelf fronts, leading to exposure to the open ocean; (III) outer margin fracturing and rifting; and (IV) initial calvings from the outer ice shelf margins along these fractures and rifts, creating long thin icebergs (termed 'sliver bergs' here), followed immediately by rapid disintegration of the adjacent interior ice shelf areas (Fig. 2; also refs ^{5,10,13,26}).

While flooding and hydrofracture (I) are the drivers of true catastrophic disintegration (not just breakup or retreat), we show that linkages between (II), (III) and (IV) have key roles in preconditioning and then triggering the disintegration of the shelves examined. Swell-induced flexure over extended periods, enhanced by the absence

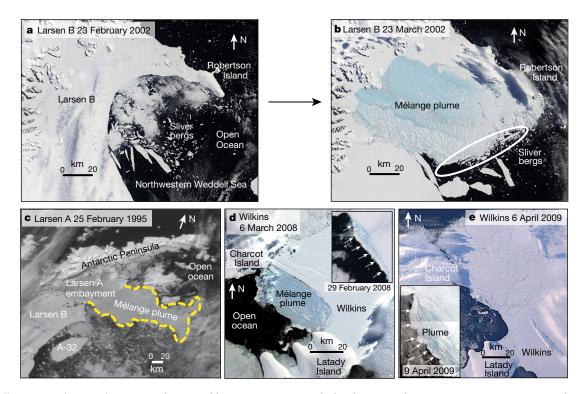


Fig. 2 | Satellite images showing the common features of four disintegration events. The common features are rapid disintegration and seaward plume dispersal immediately following outer-margin calving in the form of sliver bergs, with an absence of sea ice offshore. a, b, Larsen B 2002 (MODIS). c, Larsen A 1995 (AVHHR). d, Wilkins 2008 (MODIS). e, Wilkins 2009 (MODIS). Sliver berg calvings are labelled or (in d and e)

marked with arrows. The NOAA AVHRR image in **c** was obtained and used with permission from the British Antarctic Survey (http://www.nercbas.ac.uk/icd/bas_publ.html). Satellite images from the NASA MODIS instrument in **a**, **b**, **d** and **e** were obtained from the NASA NSIDC DAAC archive (http://nsidc.org/data/iceshelves_images/) (see Methods⁵¹).

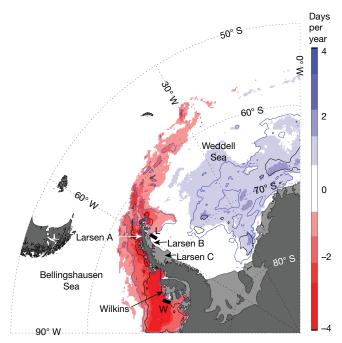


Fig. 3 | Trend map of satellite-derived annual sea ice season duration for the Weddell and Bellingshausen seas for 1979/1980 to 2009/2010. The reduction in the annual sea ice season offshore from the Wilkins and Larsen ice shelves can be seen (with red indicating fewer days of coverage, that is, shorter sea ice seasons). The black/blue contours delimit significance at P < 0.01 and 0.10, respectively (with standard error determined using the effective degrees of freedom present in the regressive residuals). The background map is based on the CIA World Map database (and we produced it using IDL; see Methods).

of a sea ice buffer, is a preconditioning factor through the process of outer-margin fracture amplification (along with the glaciological processes outlined above). Subsequent rapid fine-scale sliver-berg calving is the trigger mechanism, that is, the most proximal cause for the onset of wider disintegration (the 'straw that broke the camel's back'). This calving unleashes the stored potential energy in the system in the form of perched water in ponds and fractures, and toppling of the taller ice shelf blocks into a mélange of fragments.

Loss of a sea ice buffer

A decadal-scale reduction of sea ice coverage (concentration, that is the proportion of the ocean surface covered by sea ice, and duration) over the satellite era (since 1979) in the northwestern Weddell Sea and the Bellingshausen Sea (Fig. 3; also ref. ²⁷) dramatically increased the potential for substantial ocean wave energy to reach the ice shelf fronts in mid- to late summer and in early autumn. An increase in open-water duration of approximately three months occurred between 1979/1980 and 2009/2010²⁷.

Time series of mean daily sea ice concentrations offshore (from the boxes marked L and W in Fig. 3a) show greater prevalence and frequency of open-water/low sea ice concentration (\leq 40%) conditions after 1991 off the Larsen A and B ice shelves and after 1989 off the Wilkins Ice Shelf (Fig. 4a, b; see also ref. ²⁸). Enhanced exposure of the Larsen ice shelf fronts occurred in November–March (see seasonal and mean concentrations in Fig. 4c and Extended Data Tables 1 and 2, respectively), coinciding with seasons of strong surface melting for these shelves⁵. For the Wilkins Ice Shelf, enhanced exposure occurred later in the season, mainly in February–May, with anomalously low monthly mean concentrations (of about 66%) extending into June in 2008 and 2009 (Fig. 4d and Extended Data Table 3). Mean sea ice concentrations adjacent to the shelf-front regions (in the boxed regions L and W in Fig. 3) decreased during these seasons over 1979/1980 to 2009/2010 (Extended Data Fig. 1).

Coincidence of disintegration and exposure

Potential causality between recent ice shelf disintegration and regional sea ice loss is highlighted by the temporal coincidence that occurred during the five major events examined here (Fig. 2). The sea ice loss in each case resulted in extensive and sustained periods of exposure to broadly open-ocean conditions offshore (Fig. 4 and Extended Data Figs. 2–4). In effect, the regional sea ice decrease created greater frequency and duration of exposure of the Larsen and Wilkins Ice Shelf fronts to ocean swells via open water or corridors of low sea ice concentration connecting them to the open seas of the southern Atlantic Ocean and Bellingshausen Sea, respectively. The fact that the Wilkins Ice Shelf disintegrations took place in austral autumn (in addition to summer) in regions where sea ice coverage has decreased in summerautumn is further evidence for a potential relationship between low sea ice coverage and ice shelf collapse.

In the case of the Wilkins Ice Shelf collapse, observations dating back to 1947 show that frontal change was minimal before 1990^{5,29,30}, that is, during and before the earlier epoch of heavy and nearly year-round sea ice coverage shown in Fig. 4b. Since 1990, that is, during the subsequent epoch of greater exposure to swells, the Wilkins Ice Shelf has lost about 40% (or 5,500 km²) of its area, mainly from its more exposed northern and northwestern fronts (marked N and W in Extended Data Fig. 5a)³⁰. This includes further strong temporal coincidence between earlier breakup events in 1990/1991 (about 650 km²), 1992/1993 (550 km²) and 1998 (about 1,100 km²)³⁰ and prolonged exposure (Fig. 4b). By contrast, breakup from the more sheltered southwestern front (marked SW in Extended Data Fig. 5a and protected by Latady Island from swells encroaching from the north/northwest) has been minimal³⁰.

At the time of the Larsen B Ice Shelf collapse in 2002 and the collapse of the Wilkins Ice Shelf in 2008 and 2009, the scope of coincident sea-ice-free conditions was exceptional. The Bellingshausen Sea was almost completely sea-ice-free over extended periods in late austral summer through to early autumn in 2008 and autumn in 2009 (Fig. 4b, d; Extended Data Fig. 2c-e). For the Larsen B Ice Shelf collapse in 2002, disintegration occurred during an unusually long period of continuous open-ocean exposure, lasting approximately 4.5 months (mid-November 2001 to late March 2002; see Fig. 4a, c; Extended Data Fig. 2b). The mean austral summer (December–February) sea ice concentration offshore in 2001/2002 was only 14.7%, compared with 99.4% the previous year (Extended Data Table 2). This major exposure event was associated with sustained strong and warm northerly/ northwesterly airflow across the Antarctic Peninsula^{6,31}, and frequent but more localized föehn wind events³². The warm winds simultaneously drove the sea ice away from the Larsen B Ice Shelf front (northwestern Weddell Sea)³¹ and caused extensive surface melt-pond coverage⁶.

Increased swell-margin interaction

Daily maximum ocean-wave hindcast data offshore from the Larsen and Wilkins (box areas L and W in Fig. 3) show peak periods of 12 s and 16 s, and peak significant wave heights (that is, four times the standard deviation of the ocean surface elevation) of 3.2 m and 6.5 m, respectively, in the periods before and during the disintegrations (Extended Data Figs. 3a, b and 4a, b). Pink and blue horizontal bars in Extended Data Figs. 3 and 4 indicate the extended prevalence of on-ice shelf swells and absence of sea ice, respectively. A model of wave attenuation in moderate (50% concentration) to heavy (90% concentration) sea ice cover shows that sea ice strongly attenuates swells with periods of 5 s to 20 s, with the wave-energy e-folding distance increasing from the order of 1 km to the order of 100 km as wave period increases, depending on sea ice concentration and thickness (Fig. 5a; see also refs ^{33,34}). This suggests that, when present, sea ice buffers the shelves from swell impacts, with thick consolidated ice providing more effective damping of wave energy than a thin and more dispersed sea ice cover³⁵.

Figure 5b, c shows model predictions of the maximum flexural strains imposed on shelves by regular incident swells, as functions of wave period with and without a sea ice buffer, in which shelf thicknesses are representative of the outer Larsen A and B and Wilkins ice shelves,

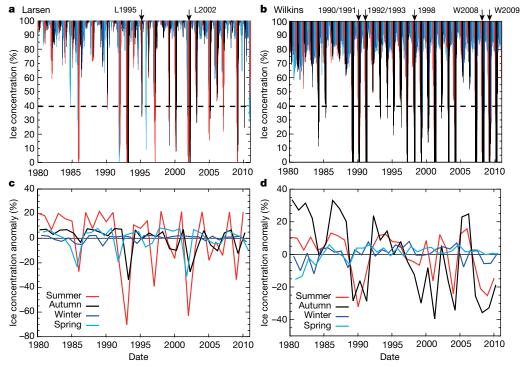


Fig. 4 | Time series showing reduction in sea ice coverage offshore from the Wilkins and Larsen ice shelves, and temporal coincidence with disintegration events. a, b, Time series of satellite passive microwave-derived daily sea ice concentrations for the regions L and W in Fig. 3 off the Larsen A and B and Wilkins ice shelves, respectively, with arrows denoting the approximate onset timings of the major disintegration events.

For Wilkins in 1998 (b), the northern part of the shelf disintegrated in March but the fragments remained in place for the subsequent decade. The plots are colour-coded to highlight summer (red), autumn (black), winter (dark blue) and spring (light blue). c, d, Plots of mean seasonal sea ice concentration anomalies off the Larsen A and B and Wilkins ice shelves, respectively, from the same regions.

respectively^{5,7}. The modelled sea ice buffer used for Fig. 5c includes an additional 50 km of consolidated landfast ice attached to the ice shelf front, to represent observed conditions in the months before the disintegration events of the Wilkins Ice Shelf in 2008 and 2009 (Extended Data Fig. 5). In addition, Fig. 5d shows modelled strains on an ice shelf 80 m thick (with wave height 3 m). This case represents locally thinner areas near the ice shelf front resulting from bottom crevassing, which is thought to be one of the weakening factors contributing to the disintegration events. This is likely to have been a component of the Larsen A in 1995 and Larsen B in 2002 ice shelf disintegration events, on the

basis of recently acquired radar profiles from the outer Larsen C (an ice shelf to the south of Larsen B that is currently undergoing similar transformations to those observed for Larsen A and B before disintegration). Basal crevasses beneath overlying surface crevasses decrease the thickness of coherent ice^{36,37} by an estimated 24%–66%.

Longer-period swells induce larger maximum strains on the shelves because they transmit a greater proportion of their energy beyond the shelf front ^{17,38,39}. At the maximum peak wave periods observed (12 s for Larsen and 16 s for Wilkins) and with a buffer of sea ice present that is 125 km wide and at 90% concentration, Larsen experiences strains

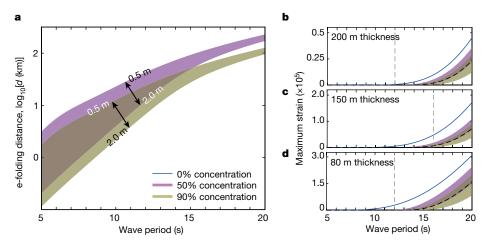


Fig. 5 | Results from a sea ice attenuation model and ice shelf-wave flexure model. a, Wave-energy e-folding distance $\log_{10}d$, where d is distance in kilometres, as a function of wave period for a sea ice zone with concentrations of 50% and 90%, for sea ice thickness ranging from 0.5 m to 2 m and mean floe length of 100 m. **b–d**, Predictions of maximum flexural strain imposed on shelves of different thickness by regular incident swells, as functions of wave period, where purple is a sea ice cover of 50%

concentration and width ranging from 80~km to 250~km, and olive is a sea ice cover of 50% concentration and width ranging from 80~km to 250~km, for a shelf 200~m thick and wave height 3~m (\mathbf{b}); a shelf 150~m thick and wave height 6~m (\mathbf{c}); and a shelf coherent thickness of 80~m and wave height 3~m (\mathbf{d}). The curved black dashed line denotes a 90% concentration sea ice cover 125~km wide. Vertical dashed lines are maximum peak wave periods observed.

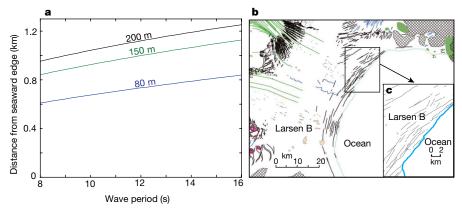


Fig. 6 | **Maximum swell-induced strain and enhanced fracture in the ice shelf front zone. a**, Model predictions of maximum strain location relative to the seaward ends of ice shelves 80 m thick, 150 m thick and 200 m thick, as a function of wave period. **b**, Structural interpretation of high-resolution satellite images of the Larsen B Ice Shelf, showing rifts and

crevasses near the ice shelf edge in November 2001, that is, 2 to 3 months before the Larsen B 2002 disintegration. **c**, Close up of the area in the box in **b**, showing rifts and crevasses near and sub-parallel to the ice shelf edge (light blue line). (Panels **b** and **c** after figures 3c and 4d in ref. ¹², with permission of the authors and Cambridge University Press.)

of the order of 10^{-12} , and Wilkins experiences a strain of the order of 10^{-7} , noting that it has an additional 50 km of landfast ice. With the sea ice buffer removed, these strains increase to approximately the order of 10^{-8} and 10^{-6} for the Larsen ice shelves and the Wilkins Ice Shelf, respectively. The increase is particularly dramatic for the Larsen ice shelves (four orders of magnitude) because the peak wave period is shorter than for the Wilkins Ice Shelf (although the order-ofmagnitude increase there is still substantial). For 80 m-thick areas of the outer Larsen Ice Shelf, thinned due to bottom crevassing^{36,37}, the strain increases by an additional two orders of magnitude to around the order of 10^{-6} . Moreover, strains on both the Larsen and Wilkins increase to around the order of 10^{-5} for slightly longer periods. Strains of this magnitude acting over the prolonged periods of sea ice absence observed are sufficient to precondition and eventually trigger sliver-berg calving (and then disintegration) by enlarging existing weaknesses in outer ice shelf margins through flexural cycling and fatigue^{14,15}.

Prior to their disintegration, the Larsen A and B Ice Shelf fronts and the north-northwestern Wilkins Ice Shelf front were all characterized by networks of long, closely spaced transverse rifts nearly parallel to the ice shelf front in a zone approximately 5–10 km wide ^{12,13,22} (Fig. 6b, c). For Larsen B, these rifts increased greatly in length and number after January 2000, that is, in the lead-up to the 2002 disintegration of the Larsen B Ice Shelf. The model outputs show that the maximum strains at the wave periods observed are attained in the outer margins of the shelves, over the first few kilometres in from their seaward edge (Fig. 6a, Extended Data Fig. 6), that is, within the zone of observed shelf-front fracture (Fig. 6b, c) and sliver iceberg formation (Fig. 2). Thinner shelves experience greater strain because they reflect less incident wave energy, and the maximum strain occurs closer to the shelf front.

In all cases examined, the extensive outer-rift networks spawned major sliver-berg calvings across wide fronts (Fig. 2; see also figure 4 of ref. ¹³) that initiated the disintegration events for the ice shelf interior areas. These events occurred during periods of sustained on-shelf swells caused by the absence of sea ice that allowed broad exposure to the Southern Ocean (Extended Data Figs. 3 and 4). The apparent lack of (exact) temporal correspondence between (1) peak wave energy and resultant strain and (2) disintegration onset in Extended Data Figs. 3 and 4 is not unexpected, given the complexity of the interactions between the ice shelf and the waves and the different timescales of the processes involved. Rather, a critical factor is that observed periods of large strains are likely to work pre-existing weaknesses in the outer ice shelf margins to the point where sliver-berg calving occurs along the fractures. High-strain cycles may also open up existing fractures to allow water in, which will reduce the fracture toughness and increase the likelihood of breakage (ice shelves with water-filled rifts are weaker than those with dry crevasses 40). As noted, basal crevasses near the

ice shelf fronts further increase the likelihood of swell-induced failure there. In the case of the collapse of Larsen A Ice Shelf in 1995, the apparent lack of sufficient strain at the time of disintegration (Extended Data Fig. 3c) also results from the fact that the disintegration event coincided with a shorter period of conditions of low sea ice concentration offshore (Fig. 2c), driven by strong northwesterly winds in late January¹³, but prolonged open-water conditions to the northeast (Extended Data Fig. 2a; see Methods).

Occurring immediately before each disintegration event, sliver-berg calvings (Fig. 2) removed keystone blocks from the arch-like configuration of the ice shelf front that were crucial to its structural integrity. As described⁴¹ for Larsen B, this concave configuration mirrored a stabilizing 'compressive arch' in the horizontal strain rate trajectories of the ice shelf (the stress field), that is, a band of transversely compressed ice in the shelf. We propose that swell-induced structural failure of the outer-shelf margin in the absence of a sea ice buffer triggered catastrophic retreat of the weakened ice shelf that had been braced by the previously intact outer ice shelf margin. This is consistent with model predictions 41 that rapid irreversible retreat would occur (to some other stable configuration) should the ice front break back only a few kilometres through the compressive arch (see also refs ^{3,29}). Moreover, it has been proposed⁴² that the long, thin and relatively stable sliver bergs trap tsunamic waves between their landward margins and the eroding ice shelf front. These are locally generated by the toppling of calved icebergs and have much larger amplitudes than do ocean swells, exacerbating break-up.

The schematic in Fig. 1 summarizes the complex interplay and sequence of the different (though common) factors involved in ice shelf preconditioning for collapse and subsequent disintegration events (for example, in Fig. 2), highlighting potential cross-cryosphere linkages and the role of sea ice loss. These common factors are underpinned by large-scale change/variability in atmospheric and ocean processes reported elsewhere, for example in ref. ²⁷. Outer-margin breakage (failure) in the absence of a sea ice buffer was the trigger that set in motion a spontaneous runaway pattern ('chain reaction') of catastrophic largescale collapse, driven by the stored potential energy of the system (reported in previous studies, such as ref. 26), of ice shelves weakened by inherent glaciological discontinuities (crevasse and suture zones) and decades of melt and thinning. Thus, water driving into cracks⁵, chain-reaction drainage of surface lakes⁴³, ice blocks toppling like dominoes²⁶, and localized tsunami-genesis⁴² combine to consume the shelf area prone to disintegration²⁶. Furthermore, the lack of sea ice offshore probably facilitated the characteristic rapid release and seaward dispersal of the plumes of thousands of small icebergs and ice mélange into the open ocean shown in Fig. 2b, c (and driven by ocean currents and the same strong northwesterly winds that removed the sea ice in the

case of the collapses of the Larsen A Ice Shelf in 1995 and of the Larsen B Ice Shelf in $2002^{8,13}$).

The additional role of landfast sea ice

In the case of the Wilkins Ice Shelf, there is also strong temporal coincidence between disintegration and the break-up or absence of consolidated landfast ice attached to the northern and northwestern ice shelf fronts that are most exposed to ocean swell (Extended Data Fig. 5c, e, f). This in turn coincides with open-ocean conditions (lack of a wider pack-ice buffer), reflecting the sensitivity of unprotected landfast ice itself to disintegration by ocean swell⁴⁴. In contrast, and in the summer of 2006, extensive landfast ice (see Extended Data Fig. 5a) coincided with persistent heavy pack ice conditions offshore and, notably, no ice shelf disintegration occurred (Fig. 4b). These observations imply that landfast ice may have an important role as an additional buffer against swell impacts, and also in mechanically bonding (buttressing) the exposed weakened outer margins of the Wilkins Ice Shelf, to maintain their structural integrity and reduce calving. This is in line with observations from the Mertz Glacier Tongue in East Antarctica⁴⁵. Landfast ice can also bond (freeze) together those fragments from disintegration events (that is, mélange and icebergs) that do not drift away, to re-form a hybrid ice shelf¹⁰.

Although loss of landfast ice appears to have played little or no part in the Larsen B Ice Shelf 2002 disintegration, ref. ¹³ speculates that a combination of landfast ice and cold temperatures held Larsen A in place in 1994, in spite of the presence of extensive ice shelf fracturing. Major disintegration then only occurred after 20 January 1995 and following landfast-ice removal, in response to strong northwesterly winds and high air temperatures. Moreover, there is an ongoing apparent calving suppression in the remnant Larsen B Ice Shelf (E. Pettit, personal communication, 2017) due to the presence of thick multi-year landfast ice in the embayment of the former ice shelf.

What does the future hold?

Accurate and realistic numerical modelling of ice shelf buttressing is a critical step towards predicting the response of the Antarctic Ice Sheet to climate change and its contribution to sea-level rise¹¹. While progress is being made towards prediction of weakening of remaining ice shelves, for example, ref. ⁴⁶, our results suggest the additional need for ice sheet models to include a modified calving parameterization that accounts for sea ice and swells. Sea ice both buffers vulnerable ice shelf frontal zones from destructive ocean swells and, in the case of landfast ice, mechanically bonds and buttresses the outer ice shelf margins. This is in addition to the need to incorporate the effects of (changing) sea ice conditions on oceanic heat input to ice shelf cavities and the subsequent effect of altered basal melt rates⁴⁷, as well as very long-period (more than 20 s) waves that are largely unaffected by the presence of sea ice ^{16,18–20} (see Methods).

The Wilkins Ice Shelf is the most southerly and largest ice shelf to be affected by disintegration events to date. South of Larsen B, the Larsen C Ice Shelf has thinned⁴⁸, with continued warming temperatures and recent large-scale calving of a giant iceberg⁴⁹, similar to the earlier trends that led to the collapse of Larsen A and B. The sequence of steps illustrated in Fig. 1, and discussed here, suggests that an approximate predictive path for ice shelf disintegration can be mapped out, with order-of-magnitude timescales keyed to observed and linked changes in ice (both glacial and sea ice), ocean and atmosphere conditions. The areas most susceptible are the northern Larsen C Ice Shelf and the remnants of the Larsen B and Wilkins ice shelves, as well as the West, Shackleton and outer Amery ice shelves in East Antarctica⁵.

A proviso is that not all remaining ice shelves will necessarily be affected in the same way by sea ice loss or change, given their different glaciological characteristics and physical settings²⁹. Rapid disintegration requires a flooded ice shelf with very low internal compression, extensive hydrofracturing and loss of remaining backstress and extensive fracturing at the ice shelf front. In the absence of this essential preconditioning, seasonal interaction of ice shelf fronts with long-period

swell (during open-water conditions at the ice shelf front) or yearround interaction with intermittent tsunami and infragravity waves produces only rare small-scale calvings²⁰. Cool-climate ice shelves that have frequent open-water conditions offshore, such as the Fimbul and Ross ice shelves, are currently stable. Recent work on the Ross Ice Shelf³⁹ suggests that flexure from waves may contribute to rifting and cracking when acting over long (multi-year to decadal) timescales. However, ice shelves generally withstand this without sudden retreat, if their surfaces are not flooded and overall ice shelf structural integrity is not compromised. Only highly confined ice shelves can tolerate frequent and extensive seasonal meltwater flooding, such as the inner Amery Ice Shelf and the George VI Ice Shelf. Moreover, the destructive effect of hydrofracture could be mitigated in ice shelves where large river networks efficiently export meltwater from the ice shelf surface to the ocean, such as in the Nansen Ice Shelf⁵⁰. Clearly, more work is required to identify, quantify and model all processes affecting ice shelf stability and their inter-relationships.

An additional unknown is the potentially destructive effect of ocean waves breaking against and undercutting the outer cliff face of an ice shelf in the absence of a sea ice buffer. By compressing air in rifts and cracks and explosively eroding the vulnerable ice shelf margin, this would also probably contribute to the mechanical failure and collapse of ice-cliff structures and the calving of keystone sliver bergs.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0212-1.

Received: 8 July 2017; Accepted: 3 April 2018; Published online 13 June 2018.

- Bindschadler, R. et al. Getting around Antarctica: new high-resolution mappings of the grounded and freely-floating boundaries of the Antarctic ice sheet for the International Polar Year. Cryosphere 5, 569–588 (2011).
- Gudmundsson, G. H. Ice-shelf buttressing and the stability of marine ice sheets. Cryosphere 7, 647–655 (2013).
- Fürst, J. J. et al. The safety band of Antarctic ice shelves. Nat. Clim. Chang. 6, 479–482 (2016).
- Morris, E. & Vaughan, D. in Antarctic Peninsula Climate Variability: Historical and Paleoenvironmental Perspective (eds Domack, E. et al.) 61–68 (American Geophysical Union, Washington DC, 2003).
- Scambos, T., Hulbe, C. & Fahnestock, M. in Antarctic Peninsula Climate Variability: Historical and Paleoenvironmental Perspective (eds Domack, E. et al.) 335–347 (American Geophysical Union, Washington DC, 2003).
- van den Broeke, M. Strong surface melting preceded collapse of Antarctic Peninsula ice shelf. Geophys. Res. Lett. 32, L12815 (2005).
- Braun, M. & Humbert, A. Récent retreat of Wilkins lce Shelf reveals new insights in ice shelf breakup mechanisms. *IEEE Geosci. Remote Sens. Lett.* 6, 263–267 (2009).
- Rack, W. & Rott, H. Pattern of retreat and disintegration of the Larsen B ice shelf, Antarctic Peninsula. Ann. Glaciol. 39, 505–510 (2004).
- Vieli, A., Payne, A. J., Shepherd, A. & Du, Z. Causes of pre-collapse changes of the Larsen B ice shelf: numerical modelling and assimilation of satellite observations. *Earth Planet. Sci. Lett.* 259, 297–306 (2007).
- Scambos, T. A. et al. Ice shelf disintegration by plate bending and hydrofracture: satellite observations and model results of the 2008 Wilkins Ice Shelf break-ups. Earth Planet. Sci. Lett. 280, 51–60 (2009).
- Khazendar, A., Rignot, E., Larour, E. & Larsen, B. Ice shelf rheology preceding its disintegration inferred by a control method. *Geophys. Res. Lett.* 34, L19503 (2007).
- Glasser, N. F. & Scambos, T. A. A structural glaciological analysis of the 2002 Larsen B ice shelf collapse. J. Glaciol. 54, 3–16 (2008).
- Rott, H., Skvarca, P. & Nagler, T. Rapid collapse of Northern Larsen Ice Shelf. Antarct. Sci. 271, 788–792 (1996).
- Holdsworth, G. & Glynn, J. Iceberg calving from floating glaciers by a vibration mechanism. Nature 274, 464–466 (1978).
- Squire, V. A., Robinson, W. H., Meylan, M. H. & Haskell, T. G. Observations of flexural waves in the Erebus Glacier Tongue, McMurdo Sound, Antarctica, and nearby sea ice. J. Glaciol. 40, 377–385 (1994).
- Sergienko, O. V. Elastic response of floating glacier ice to impact of long-period ocean waves. J. Geophys. Res. 115, F04028 (2010).
- Robinson, W. & Haskell, T. G. Travelling flexural waves in the Erebus Glacier Tongue, McMurdo Sound, Antarctica. Cold Reg. Sci. Technol. 20, 289–293 (1992).
- MacAyeal, D. R. et al. Transoceanic wave propagation links iceberg calving margins of Antarctica with storms in tropics and Northern Hemisphere. Geophys. Res. Lett. 33, L17502 (2006).

- Bromirski, P. D., Sergienko, O. V. & MacAyeal, D. R. Transoceanic infragravity waves impacting Antarctic ice shelves. Geophys. Res. Lett. 37, L02502 (2010).
- Brunt, K. M., Okal, E. A. & MacAyeal, D. R. Antarctic ice-shelf calving triggered by the Honshu (Japan) earthquake and tsunami, March 2011. J. Glaciol. 57, 785–788 (2011).
- Robinson, W. & Haskell, T. G. Calving of Erebus Glacier tongue. Nature 346, 615–616 (1990).
- 22. Braun, M., Humbert, A. & Moll, A. Changes of Wilkins Ice Shelf over the past 15 years and inferences on its stability. *Cryosphere* **3**, 41–56 (2009).
- Domack, E. et al. Stability of the Larsen B ice shelf on the Antarctic Peninsula during the Holocene epoch. Nature 436, 681–685 (2005).
- Rignot, E. et al. Accelerated ice discharge from the Antarctic Peninsula following the collapse of Larsen B Ice Shelf. Geophys. Res. Lett. 31, L18401 (2004).
- Scambos, T. A., Bohlander, J. A., Shuman, C. A. & Skvarca, P. Glacier acceleration and thinning after ice shelf collapse in the Larsen B embayment, Antarctica. Geophys. Res. Lett. (2004).
- MacAyeal, D. R., Scambos, T. A., Hulbe, C. L. & Fahnestock, M. A. Catastrophic ice shelf breakup by an ice shelf fragment capsize mechanism. J. Glaciol. 49, 22–36 (2003).
- Stammerjohn, S., Massom, R., Rind, D. & Martinson, D. Regions of rapid sea ice change: an inter-hemispheric seasonal comparison. *Geophys. Res. Lett.* 39, L06501 (2012).
- Padman, L. Oceanic controls on the mass balance of Wilkins Ice Shelf, Antarctica. J. Geophys. Res. Oceans 117, C01010 (2012).
- 29. Cook, A. J. & Vaughan, D. G. Overview of areal changes of the ice shelves on the Antarctic Peninsula over the past 50 years. *Cryosphere* **4**, 77–98 (2010).
- Arigony-Neto, J. et al. in Global Land Ice Measurements from Space (eds Kargel, J. et al.) Ch. 30, 717–741 (Springer Praxis, Heidelberg, 2014).
- Turner, J., Harangozo, S. A., Marshall, G. J., King, J. C. & Colwell, S. R. Anomalous atmospheric circulation over the Weddell Sea, Antarctica, during the austral summer of 2001/02 resulting in extreme sea-ice conditions. *Geophys. Res. Lett.* 29, 2160 (2002).
- Cape, M. R. et al. Foehn winds link climate-driven warming to ice shelf evolution in Antarctica. J. Geophys. Res. Atmos. 120, 11037–11057 (2015).
- Bennetts, L. G. & Squire, V. A. On the calculation of an attenuation coefficient for transects of ice-covered ocean. Proc. R. Soc. Lond. A 468, 136–162 (2012).
- Meylan, M. H., Bennetts, L. G. & Kohout, A. L. In situ measurements and analysis of ocean waves in the Antarctic marginal ice zone. *Geophys. Res. Lett.* 41, 5046–5051 (2014).
- Kohout, A. L. & Meylan, M. H. An elastic plate model for wave attenuation and ice floe breaking in the marginal ice zone. J. Geophys. Res. 113, C09016 (2008).
- Luckman, A. et al. Basal crevasses in Larsen C Ice Shelf and implications for their global abundance. Cryosphere 6, 113–123 (2012).
- McGrath, D. et al. Basal crevasses on the Larsen C Ice Shelf, Antarctica: implications for meltwater ponding and hydrofracture. Geophys. Res. Lett. 39, L16504 (2012).
- Fox, C. & Squiré, V. A. S. Coupling between the ocean and an ice shelf. Ann. Glaciol. 15, 101–108 (1991).
- Bromirski, P. D. et al. Ross ice shelf vibrations. *Geophys. Res. Lett.* 42, 7589–7597 (2015).
- van der Veen, C. J. Fracture propagation as means of rapidly transferring surface meltwater to the base of glaciers. Geophys. Res. Lett. 34, L01501 (2007).
- Doake, C. S. M., Corr, H. F. J., Rott, H., Skvarca, P. & Young, N. W. Breakup and conditions for stability of the northern Larsen Ice Shelf, Antarctica. *Nature* 391, 778–780 (1998).
- MacAyeal, D. R., Abbot, D. S. & Sergienko, O. V. Iceberg capsize tsunamigenesis. Ann. Glaciol. 52, 51–56 (2011).
- Banwell, A. F., MacAyeal, D. R. & Sergienko, O. V. Breakup of the Larsen B Ice Shelf triggered by chain reaction drainage of supraglacial lakes. *Geophys. Res. Lett.* 40, 5872–5876 (2013).

- Langhorne, P. J., Squire, V. A., Fox, C. & Haskell, T. G. Lifetime estimation for a land-fast ice sheet subjected to ocean swell. *Ann. Glaciol.* 33, 333–338 (2001).
- Massom, R. A. et al. Examining the interaction between multi-year landfast sea ice and the Mertz Glacier Tongue, East Antarctica: another factor in ice sheet stability? J. Geophys. Res. 115, C12027 (2010).
- Borstad, C. et al. A constitutive framework for predicting weakening and reduced buttressing of ice shelves based on observations of the progressive deterioration of the remnant Larsen B Ice Shelf. Geophys. Res. Lett. 43, 2027–2035 (2016).
- Khazendar, A. et al. Observed thinning of Totten Glacier is linked to coastal polynya variability. Nat. Commun. 4, (2013).
- Holland, P. R. et al. Oceanic and atmospheric forcing of Larsen C Ice-Shelf thinning. Cryosphere 9, 1005–1024 (2015).
- Hogg, A. E. & Hilmar Gudmundsson, G. Impacts of the Larsen-C Ice Shelf calving event. Nat. Clim. Chang. 7, 540–542 (2017).
- 50. Bell, R. et al. Antarctic ice shelf potentially stabilized by export of meltwater in surface river. *Nature* **544**, 344–348 (2017).
- Scambos, T., Bohlander, J. & Raup, B. Images of Antarctic Ice Shelves [February 2002 to April 2009]. Larsen B and Wilkins https://doi.org/10.7265/N5NC5Z4N (National Snow and Ice Data Center, Boulder, 1996).

Acknowledgements This work contributes to Australian Antarctic Science project 4116, and was supported by the Australian Government's Cooperative Research Centres Programme through the Antarctic Climate and Ecosystems Cooperative Research Centre. It also contributes to the World Climate Research Programme (WCRP) Climate and Cryosphere (CliC) project Targeted Activity "Linkages Between Cryosphere Elements". T.A.S. acknowledges NSF PLR 17-020175 and S.E.S. acknowledges NSF PLR 1440435. V.A.S. acknowledges the US Office of Naval Research Departmental Research Initiative "Sea State and Boundary Layer Physics of the Emerging Arctic Ocean" (award number N00014-131-0279) and the University of Otago. We appreciate the assistance of N. Glasser for Fig. 6b, c illustrating shelf fractures. We thank referees J. Hutchings, E. Rogers and R. Shen for their comments, which undoubtedly strengthened the paper.

Reviewer information *Nature* thanks J. Hutchings, E. Rogers, R. Shen and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions R.A.M. conceived the research, carried out the data synthesis and led the paper writing. All authors contributed to data analysis and interpretation, and writing, with each author contributing to several aspects of the manuscript and to its ideas. T.A.S. provided the ice shelf imagery and information and ice shelf expertise. L.G.B. carried out the sea ice-wave and ice shelf-wave interaction modelling and analysis, with V.A.S. providing expert input. P.R. contributed the wave data and analysis, and analysis of sea ice concentration data. S.E.S. provided analysis of change in sea ice seasonality from satellite data.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0212-1.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.A.M. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sea ice conditions and analysis. Maps and time series of sea ice concentration or extent and anomalies and trends shown in Fig. 4 and Extended Data Figs. 1–4, the map of trends in annual sea ice duration in Fig. 3, and the mean sea ice concentration values in Extended Data Tables 1–3 were computed from daily estimates of satellite-derived sea ice concentration (at a spatial resolution of 25 km × 25 km) derived by the NASA Bootstrap algorithm⁵² for the period 1979–2010. Concentration refers here to the percentage of a given area of ocean surface covered by sea ice. These data were obtained from the NASA National Snow and Ice Data Center (NSIDC) Distributed Active Archive Center (DAAC) dataset at http://nsidc.org/data/NSIDC-0079. Anomaly values were computed relative to the long-term mean for 1979–2010. Detailed analyses are carried out on regions immediately offshore from the Larsen and Wilkins ice shelves, within the boxed regions marked L and W, respectively, in Fig. 3. The locations of the boxed regions are (1) Larsen, 65°-66° S, 58°-59° W; and (2) Wilkins 70°-71° S, 76.5°-77.5° W.

The background maps of Antarctica and ice shelves in Fig. 3 and Extended Data Fig. 2 were produced in IDL using the function 'map_continents' and option 'high resolution'. This uses the CIA World Map database (circa 1993), with data points approximately 1 km apart.

Monthly, seasonal and annual mean sea ice concentration values for the box regions offshore are given for both the earlier 'ice-covered' and later 'sea ice loss' epochs (see main text) in Extended Data Tables 1 and 2, respectively, for Larsen A and B, and for the period 1979–2010 for Wilkins in Extended Data Table 3. Mean sea ice concentrations of less than 50% are highlighted to show the increase in frequency of occurrences of low sea ice concentration in the later versus earlier epochs. Mean seasonal values are plotted for both regions (boxes) for 1979–2010 in Fig. 4c, d. These show the predominance of enhanced low-concentration/open-water periods in the austral summer off Larsen, compared to the additional signal in austral autumn to early winter (as well as summer) off Wilkins. Monthly–seasonal mean values do not, however, resolve the periods of open-water conditions that occurred offshore from the ice shelves and lasted for days to weeks, as shown in Fig. 4a, b and Extended Data Figs. 3 and 4. This is particularly the case for the Larsen A disintegration event in 1995.

We note also that mean sea ice concentration values offshore from Larsen A and B for January–February 1995 (given in Extended Data Table 2) are an overestimate of actual conditions, owing to the spurious interpretation by the sea ice concentration algorithm of the iceberg–mélange disintegration plume from the Larsen A Ice Shelf disintegration event in 1995 as sea ice (artefact shown within the circle in Extended Data Fig. 2a). The actual lack of sea ice to the northeast of Larsen A (in the prevailing direction of incoming swells) at the time of disintegration is shown in the NOAA Advanced Very High Resolution (AVHRR) satellite visible image from 25 February in Fig. 2c (image from British Antarctic Survey, http://www.nerc-bas.ac.uk/icd/bas_publ.html).

The map of the pattern of trends in annual sea ice duration for 1979/1980 to 2009/2010 (Fig. 3) was computed from the daily satellite passive microwavederived ice concentration data following ref. 53 and after ref. 54, by flagging the timings of annual ice-edge advance and retreat (demarcated by the 15% ice concentration isoline, following standard protocol used in the literature) within an annual search window. This covers the sea ice annual cycle, which runs from the mean summer minimum extent in mid-February to that in the following year, that is, day 46-410, or day 46-411 in leap years. Within each sea ice year, the annual day of advance on a pixel-by-pixel basis is taken to be the time when ice concentration in a given pixel first exceeds 15% for at least 5 days, while the day of retreat is the time when the ice concentration attains a value of less than or equal to 15% and remains that way until the end of the given sea ice year. Ice season duration is then the period between day of advance and retreat. For regions where ice survives the summer melt (and remains for multiple years), the days of advance and retreat were set to the lower and upper limits, respectively, that is, day 46 and day 410/11. Also, isolated days of missing data were interpolated from adjoining days.

Information on the distribution and extent of landfast ice attached to the Wilkins Ice Shelf fronts was obtained from cloud-free NASA Terra and Aqua satellite Moderate resolution Imaging Spectroradiometer (MODIS) imagery (spatial resolution 0.25 km for visible to 1.0 km for thermal infrared imagery), with analysis using ENVI software. Thermal infrared imagery was used at times of seasonal polar darkness. These images were obtained from the NASA NSIDC DAAC archive at http://nsidc.org/data/iceshelves_images/ 51 .

Sea ice parameter values incorporated in the theoretical modelling (see detailed description below) are approximations of mean conditions based on both the satellite sea ice concentration data and in situ and remote sensing observations from the literature. Ice concentrations are set at 0%, 50% and 90% to cover the range of values observed, for example, in Extended Data Tables 1–3 and Fig. 4. For the northwestern Weddell Sea, we parameterize mean sea ice thickness as 2 m (after in situ measurements of ice turned over by ships compiled by ref. ⁵⁵), although actual thickness may be greater. This region is characteristically covered by compact and

highly deformed multi-year sea ice that is convergent against the eastern Antarctic Peninsula in the western limb of the Weddell Gyre 56 , limiting the availability of in situ observations. Airborne electromagnetic induction measurements off Larsen A in October 2006 showed a nearshore band of deformed first-year ice with a mean thickness of about 2 m, merging into extensive coverage of heavily deformed and compact second-year ice with a mean thickness of more than 3 m (ref. 57). Following refs 58,59 , the mean floe length in the Bellingshausen and northwestern Weddell seas is taken to be 100 m.

Sea ice thickness is also set at 2 m off the Wilkins Ice Shelf. Mean sea ice measurements acquired in the Bellingshausen Sea in summer 60 had a mean value of 1.3 m, but springtime measurements of mean sea ice draft (an approximation of total ice thickness) for four floes by an autonomous underwater vehicle ranged 61 from 2.25 m to 5.48 m. As such, 2 m may again be a conservative estimate in the model. The width of consolidated landfast ice extending seawards of the Wilkins Ice Shelf is taken to be 50 km (based on imagery in Extended Data Fig. 5); the thickness of this ice is unknown (no in situ measurements are available), but is again taken to be 2 m (based on the annual maximum thickness of annual landfast ice elsewhere in Antarctica 62).

Values of sea ice zone width used in the theoretical models are derived from the mean monthly or multi-monthly maps of satellite sea ice concentration for the periods containing the disintegration events (Extended Data Fig. 2). They are based on the nearest distance from the ice shelf seaward front to the open ocean in the approximate dominant direction of swells at the time of disintegration, with the ice edge being delineated as the 15% ice concentration isoline. These distances are about 200 km for the Larsen A Ice Shelf collapse (in January 1995), about 250 km for the Larsen B Ice Shelf collapse (in January-March 2002), and about 80 km for each of the disintegrations of the Wilkins Ice Shelf (in February-March 2008, May-July 2008 and March-April 2009).

The timing and extent of ice shelf disintegration. Information on timings of the disintegration events is based on: (1) analysis of NASA MODIS and NOAA AVHRR visible and thermal infrared satellite imagery obtained from the NASA NSIDC DAAC archive at http://nsidc.org/data/iceshelves_images/⁵¹; (2) information from https://www.sciencedaily.com/releases/2008/07/080710115142.htm; and (3) the literature cited in the references, for example, ref. ³⁰. Disintegration events for the Wilkins Ice Shelf are 28 February to 6 March 2008, 27 to 31 May 2008 and 28 June to mid-July 2008¹⁰. In 2009, the narrow bridge of remaining ice shelf connecting Charcot Island to Latady Island shattered between 31 March and 6 April (https://nsidc.org/news/newsroom/20090408_Wilkins.html; see also Extended Data Fig. 5f).

Ocean wave data and analysis. Ocean wave-field data were obtained from the CAWCR (Collaboration for Australian Weather and Climate Research) Wave Hindcast 1979–2010 dataset run on a $0.4^{\circ} \times 0.4^{\circ}$ global grid $^{63,64}.$ This uses the WAVEWATCH III v4.08 wave model forced with NCEP Climate Forecast System Reanalysis (CFSR) surface winds at 0.3° spatial and hourly temporal resolutions, with the sea ice edge defined using hourly sea ice concentrations from the CFSR $dataset. \ The \ data \ were \ acquired \ from \ https://data.csiro.au/dap/landingpage?pid=c-landingpage.$ siro:6616, and analysed for the box regions offshore from the Larsen and Wilkins ice shelves and marked L and W, respectively, in Fig. 3. Time series of median values of peak wave period, significant wave height and wave direction in the lead-up to and during the disintegration events are shown in Extended Data Figs. 3 and 4. Originally defined as the mean height (from trough to crest) of the highest third of waves, significant wave height is now evaluated as four times the square root of the zeroth moment of the energy density spectrum. Regarding direction, values from the direction of the corridors of open ocean observed in the satellite-derived sea ice concentration maps, that is, 30° E to 120° E off the Larsen ice shelves and 0° to 90° W off the Wilkins Ice Shelf, are marked as pink horizontal bars to highlight the contribution of swells from these directions to ice shelf margin flexure.

Model of swell attenuation by sea ice. The predictions of swell attenuation caused by a sea ice cover shown in Fig. 5 are based on theoretical models of attenuation due to scattering and viscosity. The scattering component uses the semi-analytical approximation of ref. 65 for attenuation predicted by a model in which linear potential-flow theory models the water motions and the sea ice cover is modelled as a vast collection of floating thin-elastic plates, where floe lengths are randomly chosen around a mean value of 100 m. The viscosity component uses the Robinson–Palmer ice-viscosity model 66,67 , in which the viscosity parameter $\Gamma=13.5$ Pa s m $^{-1}$ is set to match the long-period regime of the empirical model of wave attenuation of ref. 34 in the Antarctic marginal ice zone, noting its similarity to the value $\Gamma=13$ Pa s m $^{-1}$ that refs 66,67 derived from measurements of wave attenuation in the Arctic marginal ice zone. In Fig. 5a, wave-energy e-folding distance is the distance that a swell will travel through sea ice before its energy is attenuated by a factor of $1/e \approx 37\%$.

Ice shelf and ocean wave model. The predictions of swell-induced ice shelf strain shown in Fig. 5b–d and Extended Data Figs. 3a, d, 4c, d and 6 are based on the theoretical model of ref. ⁶⁸, in which a regular swell, at a specified period and height,

is incident on an ice shelf modelled as a floating Kirchoff-Love plate. In this case, it has constant thickness and an effective Young's modulus of 11 GPa (following ref. ⁶⁹), noting that the flexural rigidity of the shelf is proportional to the Young's modulus and the thickness cubed (that is, the rigidity is far more sensitive to proportional changes in thickness than it is to the Young's modulus). Predictions are made with and without any sea ice cover present, and for different sea ice zone widths and properties based on observational data presented above. The values or range of ice shelf thickness used (80 m, 150 m and 200 m) are approximations derived from the literature^{5,22}, with 80 m included to accommodate the combined effects of surface and the likely basal crevassing in decreasing coherent ice shelf thickness towards its seaward margins (after refs ^{36,37}). The additional swell attenuation by landfast ice included in Fig. 5c is based on landfast-ice reflection predicted by the model of ref. ⁷⁰ combined with the Robinson–Palmer viscosity model (see above). Sensitivity studies (not shown) indicate that a more sophisticated model with varying shelf thickness has an insubstantial effect on the results shown in Fig. 5b-d, because the maximum strain is attained close to the seaward edge of the ice shelf (see Fig. 6a and Extended Data Fig. 6).

Predictions in Fig. 5b–d are given for both unprotected shelves (0% sea ice concentration, blue line) and shelves with a sea ice buffer of 2 m thickness at 50% concentration (purple) and 90% concentration (olive) over sea ice extents ranging from 80 km to 250 km (upper and lower bounds, respectively). Figure 5c includes an additional 50-km-wide buffer of landfast ice. The dashed profiles denote maximum strain for a 125-km-wide, 90% concentration sea ice buffer, and the vertical dashed lines denote maximum peak wave periods observed.

We recognize that the Kirchhoff–Love model is an approximation to the actual behaviour of an ice shelf in flexure, and that alternatives such as a Mindlin plate⁷¹ may be more representative of thicker ice shelves, particularly at shorter wave periods. The work of ref. ³⁸ reassures us that differences emerging from the two approaches are unlikely to modify the qualitative arguments being advanced in this paper, which are focused primarily on a hitherto-neglected aspect of ice shelf disintegration.

Very-long-period waves and tides. Very-long-period ocean infragravity and tsunami waves are not included in this analysis as they are unresolved in the wave model reanalysis data used. Moreover and to our knowledge, no contemporary observational data are available regarding ice shelf flexure and vibration (for example, from seismometer stations) in the periods leading up to and containing the disintegration events analysed in this paper. Infragravity waves are generated by the interaction of shorter-period swell from far-distant storms with coastlines⁷², and propagate vast distances across the Southern Ocean to flex the floating margins of the Antarctic Ice Sheet 18,19,73,74. Unlike the shorter-period swells examined here, infragravity waves (and tsunamis²⁰) are largely unaffected by the presence or absence of a sea ice cover^{69,75}. While the current study focuses on ocean swell with periods of less than 20 s, recent measurements from broadband seismic stations on the Ross ice shelf³⁹ show how ocean gravity waves across a range of spectra cause ice shelves about 300 m thick to vibrate up to 100 km in from their seaward front, potentially compromising their structural integrity further. Infragravity waves have also been implicated in the disintegration of part of the Wilkins Ice Shelf in 2008¹⁹.

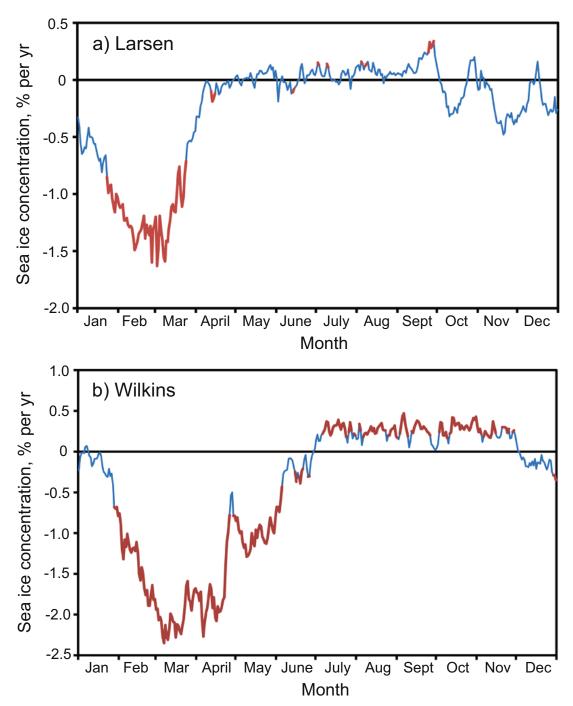
Another factor affecting the structural integrity of floating ice shelves and glacier tongues and also outside the scope of this study is tides^{76–78}. Forces exerted by tides were shown to affect rift opening tens of kilometres in from the seaward margin of the Mertz Glacier Tongue in East Antarctica⁷⁹. Other research has shown how tidal currents also induce vibrations in the floating glacier tongue⁸⁰ and small lateral pressure⁸¹, potentially causing fatigue over time. It is noted that the maximum impact of tides on ice shelf flexure that contributes to crevasse formation occurs in the grounding line owing to the long tidal wavelength^{16,78}, whereas the focus of this study is on swells as they affect the ice shelf front region.

Code availability. Analytical scripts used in this study are freely available from the authors via the corresponding author upon reasonable request.

Data availability. The datasets and products generated during this study are available from the corresponding author on reasonable request. The datasets analysed during this study are available as follows:

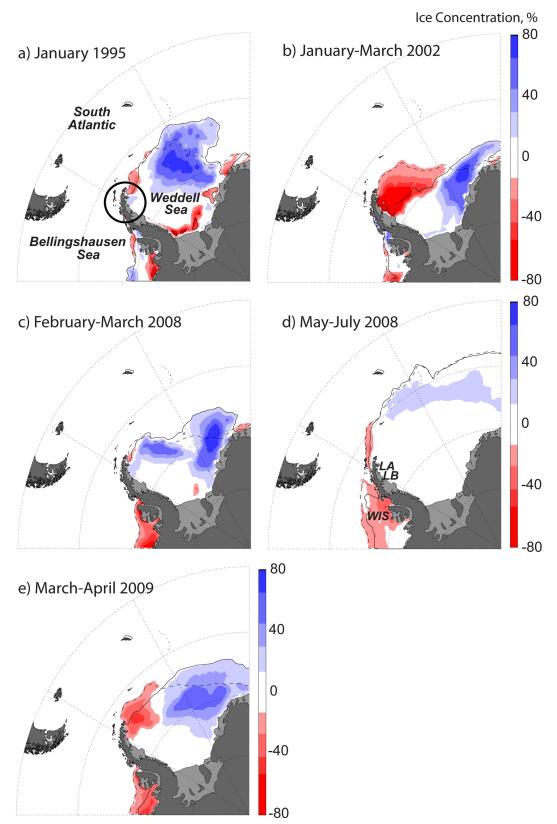
- (1) Sea ice. Daily estimates of satellite-derived sea ice concentration (gridded at a spatial resolution of 25 km \times 25 km) derived by the NASA Bootstrap algorithm for the period 1979–2010 were obtained from the NASA National Snow and Ice Data Center (NSIDC) DAAC dataset at http://nsidc.org/data/NSIDC-0079 (accessed August 2015).
- (2) Waves. Ocean wave-field data were obtained from the CAWCR (Collaboration for Australian Weather and Climate Research) Wave Hindcast 1979–2010 dataset run on a $0.4^{\circ} \times 0.4^{\circ}$ global grid at https://doi.org/10.4225/08/5 23168703DCC5 (accessed September 2017).

- (3) Satellite visible and thermal infrared imagery of ice shelves and disintegration events. The NOAA AVHRR image of the Larsen ice shelves disintegration in 1995 used in Fig. 2c was obtained from the British Antarctic Survey at http://www.nerc-bas.ac.uk/icd/bas_publ.html (accessed June 2015). The NASA MODIS images used in Fig. 2 and Extended Data Fig. 5 were obtained from the NASA NSIDC DAAC archive at http://nsidc.org/data/iceshelves_images/ (accessed June 2012).
- Comiso, J. Bootstrap Sea Ice Concentrations from NIMBUS-7 SMMR and DMSP SSM/I-SSMIS Version 2 (1979–2010) http://nsidc.org/data/NSIDC-0079 (National Snow and Ice Data Center, Boulder, 2000, updated 2015).
- Stammerjohn, S. E., Martinson, D. G., Smith, R. C., Yuan, X. & Rind, D. Trends in Antarctic annual sea ice retreat and advance and their relation to El Niño– Southern Oscillation and Southern Annular Mode variability. *J. Geophys. Res.* 113, C03S90 (2008).
- Parkinson, C. L. Trends in the length of the Southern Ocean sea-ice season, 1979–99. Ann. Glaciol. 34, 435–440 (2002).
- Worby, A. P. et al. Thickness distribution of Antarctic sea ice. J. Geophys. Res. 113, C05S92 (2008).
- Lange, M. A. & Eicken, H. The sea ice thickness distribution in the northwestern Weddell Sea. J. Geophys. Res. 96, 4821–4837 (1991).
- Haas, C. in Sea Ice (eds Thomas, D. N. & Dieckmann, G. S.) 2nd edn, 113–152 (Wiley-Blackwell, Chichester, 2010).
- Haas, C. & Viehoff, T. Sea Ice Conditions in the Bellingshausen-Amundsen Sea: Shipboard Observations and Satellite Imagery During ANT XIr3. Internal Report 51 (Department of Physics, Alfred Wegener Institute, Bremerhaven, 1994).
- Steer, A., Worby, A. & Heil, P. Observed changes in sea-ice floe size distribution during early summer in the western Weddell Sea. *Deep-Sea Res. II* 55, 933–942 (2008).
- Haas, C. Evaluation of ship-based electromagnetic-inductive thickness measurements of summer sea-ice in the Bellingshausen and Amundsen Seas, Antarctica. Cold Reg. Sci. Technol. 27, 1–16 (1998).
- 61. Williams, G. et al. Thick and deformed Antarctic sea ice mapped with autonomous underwater vehicles. *Nat. Geosci.* **8**, 61–67 (2015).
- Heil, P. Atmospheric conditions and fast ice at Davis, East Antarctica: a case study. J. Geophys. Res. 111, C05009 (2006).
- Durrant, T., Hemer, M., Trenham, C. & Greenslade, D. CAWCR Wave Hindcast 1979-2010, Version 7 Data Collection https://doi.org/10.4225/08/523168703 DCC5 (The Centre for Australian Weather and Climate Research, 2013).
- Durrant, T., Greenslade, D., Hemer, M. & Trenham, C. A Global Wave Hindcast Focussed on the Central and South Pacific. CAWCR Technical Report 070, http://www.cawcr.gov.au/technical-reports/CTR_070.pdf (The Centre for Australian Weather and Climate Research, 2014).
- 65. Bennetts, L. G. & Squire, V. A. Model sensitivity analysis of scattering-induced attenuation of ice-coupled waves. *Ocean Model.* **45–46**, 1–13 (2012).
- Williams, T. D., Bennetts, L. G., Dumont, D., Squire, V. A. & Bertino, L. Wave-ice interactions in the marginal ice zone. Part 1: Theoretical foundations. *Ocean Model.* 71, 81–91 (2013).
- Williams, T. D., Bennetts, L. G., Dumont, D., Squire, V. A. & Bertino, L. Wave-ice interactions in the marginal ice zone. Part 2: Numerical implementation and sensitivity studies along 1D transects of the ocean surface. *Ocean Model.* 71, 92–101 (2013).
- Williams, T. D. & Squire, V. A. Wave scattering at the sea-ice/ice-shelf transition with other applications. SIAM J. Appl. Math. 67, 938–959 (2007).
- Bromirski, P. D. & Stephen, R. A. Response of the Ross Ice Shelf, Antarctica, to ocean gravity-wave forcing. Ann. Glaciol. 53, 163–172 (2012).
- Fox, C. & Squire, V. A. On the oblique reflexion and transmission of ocean waves at shore fast sea ice. *Phil. Trans. R. Soc. Lond. A* 347, 185–218 (1994).
- 71. Mindlin, R. D. Influence of rotary inertia and shear on flexural motions of isotropic, elastic plates. *J. Appl. Math.* **18**, 31–38 (1951).
- Webb, S. C., Zhang, X. & Crawford, W. Infragravity waves in the deep ocean. J. Geophys. Res. 96 (C2), 2723–2736 (1991).
- Okal, E. A. & MacAyeal, D. R. Seismic recording on drifting icebergs: catching seismic waves, tsunamis and storms from Sumatra and elsewhere. Seismol. Res. Lett. 77, 659–671 (2006).
- Cathles, L. M., Okal, E. A. & MacAyeal, D. R. Seismic observations of sea swell on the floating Ross Ice Shelf, Antarctica. J. Geophys. Res. 114, F02015 (2009).
- Wadhams, P. & Doble, M. J. Sea ice thickness measurement using episodic infragravity waves from distant storms. *Cold Reg. Sci. Technol.* 56, 98–101 (2009).
- Holdsworth, G. in Oceanology of the Antarctic Continental Shelf (ed. Jacobs, S. S.) 253–271 (American Geophysical Union, Washington DC, 1985).
- Vaughan, D. G. Tidal flexure at ice shelf margins. J. Geophys. Res. 100, 6213–6224 (1995).
- 78. Padman, L., Siegfried, M. R. & Fricker, H. A. Ocean tide influences on the Antarctic and Greenland ice shelves. *Rev. Geophys.* (2018).
- Lescarmontier, L. et al. Rifting processes and ice-flow modulation observed on Mertz Glacier, East Antarctica. J. Glaciol. 61, 1183–1193 (2015).
- Lescarmontier, L. et al. Vibrations of Mertz Glacier ice tongue, East Antarctica. J. Glaciol. 58, 665–676 (2012).
- 81. Legrésy, B., Wendt, A., Tabaccó, I., Remy, F. & Dietrich, R. Influence of tides and tidal current on Mertz Glacier, Antarctica. *J. Glaciol.* **50**, 427–435 (2004).



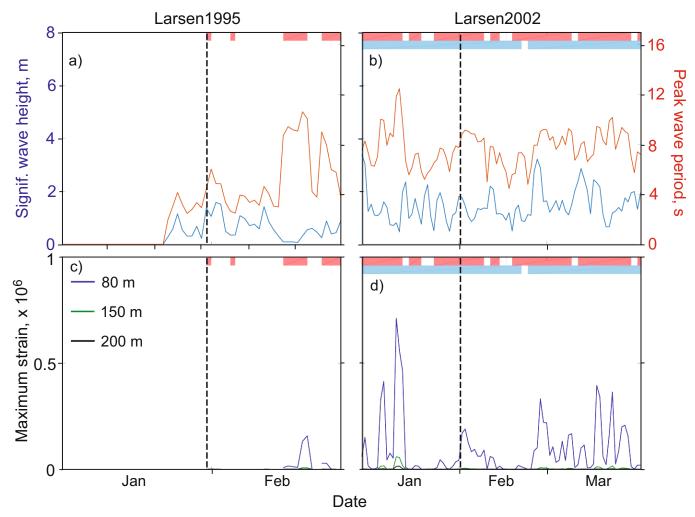
Extended Data Fig. 1 | Trends in satellite-derived daily sea ice concentration offshore of the ice shelves for 1979–2010. Data for the Larsen A and B (a) and Wilkins (b) ice shelves are from the boxes marked

L and W, respectively, in Fig. 3. Red denotes statistical significance at 90% level, while blue is not statistically significant.



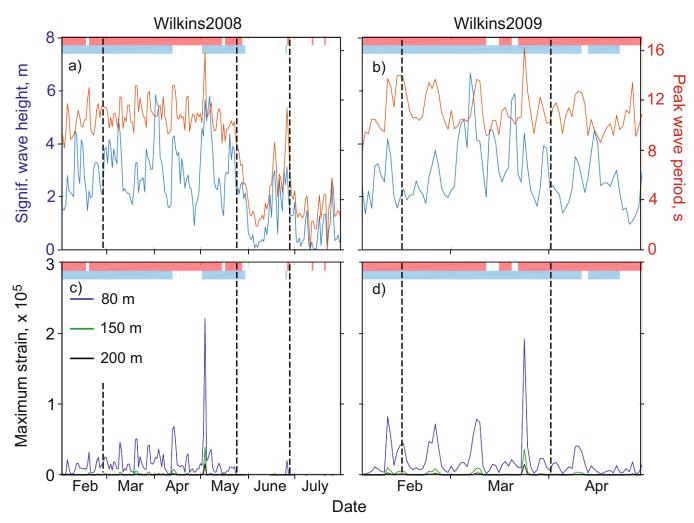
Extended Data Fig. 2 | Maps of sea ice concentration anomaly conditions during the five disintegration events. a, January 1995 (Larsen A); b, January–March 2002 (Larsen B); c, February–March 2008 (Wilkins); d, May–July 2008 (Wilkins); and e, March–April 2009 (Wilkins), versus the 1979–2010 mean for those months. The solid black contour demarcates the contemporary sea ice edge, while the dashed black contour

demarcates the mean climatological (1981–2010) ice edge location for the periods in question. See Methods for an explanation of the sea ice artefact within the black circle in **a**. The Larsen A, Larsen B and Wilkins ice shelves are marked as LA, LB and WIS in **d**. The background map is based on the CIA World Map database (and we produced it using IDL; see Methods).



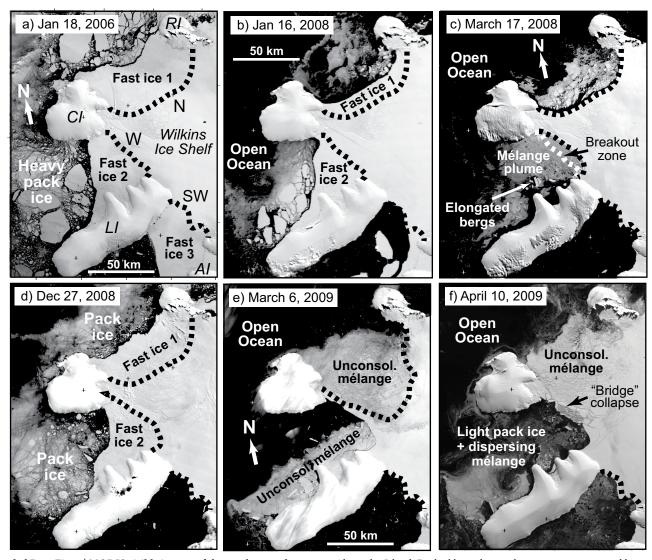
Extended Data Fig. 3 | Time series of observed wave height and peak wave period and modelled maximum ice shelf strain, in the lead up to and during the Larsen disintegration events. a, b, Daily significant wave height (blue) and peak wave period (red) within the Larsen boxed region (marked L in Fig. 3 for the Larsen A Ice Shelf collapse in 1995 and the Larsen B Ice Shelf collapse in 2002, respectively. c, d, Corresponding

model predictions of maximum ice shelf strain for an ice shelf of thickness 80 m, 150 m and 200 m. Pink horizontal bars indicate periods when waves were propagating towards the shelf from the sector 30°–120° E, and light blue bars indicate periods when the ice concentration was less than 40%. Approximate timings of disintegration event onsets are marked as vertical dashed lines.



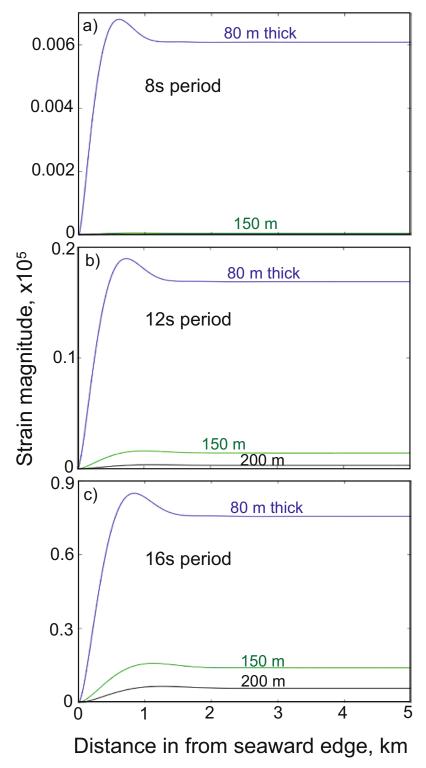
Extended Data Fig. 4 | Time series of observed wave height and peak period and modelled maximum ice shelf strain, in the lead up to and during the Wilkins disintegration events. a, b, Daily significant wave height (blue) and peak wave period (red) within the Wilkins boxed region (marked W in Fig. 3 for the Wilkins Ice Shelf disintegration events in 2008 and 2009, respectively. c, d, Corresponding model predictions of

maximum ice shelf strain for an ice shelf of thickness $80\,\mathrm{m}$, $150\,\mathrm{m}$ and $200\,\mathrm{m}$. Pink horizontal bars indicate periods when waves were propagating towards the shelf from the sector 0° – 90° W, and light blue bars indicate periods when the sea ice concentration was less than 40%. Approximate timings of disintegration event onsets are marked as dashed lines.



Extended Data Fig. 5 | MODIS visible images of the northern and western boundaries of the Wilkins Ice Shelf showing the presence or absence of landfast ice. a, 18 January 2006; b, 16 January 2008; c, 17 March 2008; d, 27 December 2008; e, 6 March 2009; and f, 10 April 2009. CI is Charcot Island, LI is Latady Island, RI is Rothschild Island and AI is

Alexander Island. Dashed lines denote the approximate seaward limit of the ice shelf. Other features (such as 'unconsolidated mélange') marked are explained in the text. Imagery from the NASA MODIS instrument was obtained from the NASA NSIDC DAAC archive (http://nsidc.org/data/iceshelves_images/)⁵¹.



Extended Data Fig. 6 | Modelled strain magnitude as a function of distance in from the seaward ice shelf edge. Modelled for an ice shelf of thickness 80 m, 150 m and 200 m and for wave periods 8 s (a), 12 s (b) and 16 s (c). Wave height is 2 m and a regular incident swell is assumed.



Extended Data Table 1 | Monthly, annual and seasonal mean satellite-derived sea ice concentrations for the earlier 'ice-covered epoch' from the region offshore of the Larsen A and B ice shelves

	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Jan	75.5	96.8	95.2	82.5	97.5	98.0	92.3	49.2	97.8	87.9	99.5	98.3
Feb	76.9	98.3	95.2	85.7	97.9	96.6	87.1	96.5	99.1	89.7	99.7	75.3
March	93.1	98.7	99.6	94.2	97.3	97.3	94.7	90.6	99.0	95.8	97.0	82.1
April	94.4	98.6	98.9	97.7	98.6	99.3	93.6	93.3	99.8	100.0	100.0	99.5
May	94.9	98.7	99.1	93.0	99.6	99.9	99.6	81.9	99.9	99.9	92.2	100.0
June	92.5	100.0	99.5	94.4	99.2	99.9	99.5	90.4	99.9	100.0	91.7	97.5
July	97.5	100.0	99.5	98.6	99.3	99.6	88.8	91.5	100.0	100.0	89.2	100.0
Aug	96.1	100.0	99.8	97.0	87.3	94.8	91.7	95.8	99.6	100.0	93.5	100.0
Sept	95.8	99.8	99.6	93.6	94.6	69.4	89.2	95.1	96.6	99.8	92.0	99.4
Oct	97.8	94.8	99.5	94.2	93.2	90.5	79.8	95.6	99.5	99.5	92.1	99.8
Nov	98.3	95.1	95.0	95.5	88.2	94.6	36.1	96.5	92.0	99.1	96.5	99.9
Dec	99.0	98.1	86.2	98.2	86.5	95.8	6.3	99.2	79.3	99.7	99.7	99.8
Year	92.7	98.2	97.3	93.7	94.9	94.6	79.9	89.6	96.9	97.6	95.3	96.0
DJF	76.2	98.0	96.2	84.8	97.9	93.7	91.7	50.7	98.7	85.6	99.6	91.1
MAM	94.1	98.7	99.2	95.0	98.5	98.8	96.0	88.6	99.6	98.6	96.4	93.9
JJA	95.4	100.0	99.6	96.7	95.3	98.1	93.3	92.6	99.8	100.0	91.5	99.2
SON	97.3	96.6	98.0	94.4	92.0	84.8	68.4	95.7	96.0	99.5	93.5	99.7

(Box L in Fig. 3.) Mean ice concentrations of less than 50% are in boldface and boxed. The seasonal and annual data are plotted in Fig. 4c.

Extended Data Table 2 | Monthly, annual and seasonal mean satellite-derived sea ice concentrations for the later 'sea ice loss epoch' from the region offshore of the Larsen A and B ice shelves

	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Jan	96.8	43.2	2.0	89.5	95.2	96.8	32.3	99.5	79.8	65.3	99.6	11.5
Feb	99.7	92.0	0.0	96.8	64.5	80.0	30.3	99.8	43.6	24.3	98.7	30.6
March	99.8	99.1	0.1	98.0	74.4	88.7	84.2	99.5	57.8	49.6	97.7	8.6
April	99.8	98.8	76.7	98.7	99.9	100.0	99.7	96.8	95.2	97.2	99.9	85.6
May	99.7	99.6	97.7	99.4	99.7	99.6	95.3	95.2	97.1	99.4	99.7	99.5
June	99.7	99.5	99.8	100.0	99.5	100.0	98.5	97.8	99.1	100.0	100.0	99.0
July	99.0	99.3	99.8	91.3	100.0	99.9	99.9	97.0	100.0	99.4	100.0	99.6
Aug	97.8	100.0	96.5	89.3	100.0	100.0	99.9	98.5	97.9	99.9	98.7	98.4
Sept	91.9	93.9	97.5	92.1	99.5	99.8	100.0	99.0	99.4	100.0	89.1	99.5
Oct	97.8	98.0	97.3	100.0	59.8	99.0	100.0	99.2	96.1	99.5	63.4	99.1
Nov	26.8	90.3	92.5	98.5	94.0	64.2	99.8	98.9	96.4	98.5	29.1	97.2
Dec	23.2	20.6	79.3	97.6	98.9	53.7	99.8	99.4	94.5	99.9	1.9	94.9
Year	86.0	86.2	69.9	95.9	90.5	90.1	86.6	98.4	88.1	86.1	81.5	77.0
DJF	98.8	52.8	7.5	88.5	85.8	91.9	38.8	99.7	74.3	61.4	99.4	14.7
MAM	99.8	99.2	58.2	98.7	91.3	96.1	93.1	97.2	83.4	82.1	99.1	64.6
JJA	98.8	99.6	98.7	93.5	99.8	100.0	99.4	97.8	99.0	99.8	99.6	99.0
SON	72.2	94.1	95.8	96.9	84.4	87.7	99.9	99.0	97.3	99.3	60.5	98.6

(Box L in Fig. 3.) Mean ice concentrations of less than 50% are in boldface and boxed. The seasonal data are plotted in Fig. 4c.



Extended Data Table 3 | Monthly, annual and seasonal mean satellite-derived sea ice concentrations from the region offshore of the Wilkins Ice Shelf

	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Jan	88.5	87.4	84.6	77.2	85.7	84.0	79.8	89.4	93.0	85.1	77.4	44.6	70.0	84.2	89.0	90.7	89.4
Feb	80.8	87.2	88.6	76.1	89.6	71.0	81.0	85.3	85.5	87.1	38.5	9.7	32.1	75.0	88.0	86.9	78.8
March	79.6	87.4	78.1	89.3	80.7	50.4	66.3	92.5	84.8	81.1	3.8	0.9	7.9	71.1	59.2	77.9	40.4
April	78.1	92.5	84.7	87.3	79.0	41.8	49.6	86.8	84.7	69.5	32.3	35.9	1.8	84.6	52.5	59.7	59.3
May	76.6	88.4	86.3	85.4	74.4	72.0	72.1	87.7	82.5	77.3	45.8	82.3	71.7	83.1	82.3	73.0	67.8
June	73.4	93.0	92.1	87.7	76.9	92.8	91.7	89.8	86.6	76.9	87.5	83.8	94.7	85.6	93.4	94.7	92.4
July	68.6	85.4	70.5	89.9	73.8	92.9	94.0	90.9	93.1	81.0	92.7	88.4	97.3	89.6	91.8	86.6	94.6
Aug	76.8	84.6	73.5	90.8	79.6	89.3	92.8	89.5	92.4	83.6	94.2	93.0	97.0	87.7	89.4	87.8	96.5
Sept	71.7	73.5	79.3	89.0	82.1	91.0	92.6	91.9	90.4	84.9	95.5	95.4	94.5	91.6	91.5	87.8	95.4
Oct	74.7	74.8	72.8	88.4	84.5	87.6	93.5	89.7	85.5	80.9	92.7	92.8	97.7	87.3	95.8	90.7	92.4
Nov	80.8	73.0	75.2	83.5	81.0	88.9	99.2	90.0	88.2	83.0	90.9	90.9	93.7	92.9	92.0	86.3	91.2
Dec	90.6	90.5	88.5	90.4	86.6	86.5	98.4	89.7	88.1	90.2	83.1	87.4	85.0	91.4	93.6	89.4	89.3
Year	78.4	84.8	81.2	86.3	81.2	79.0	84.3	89.4	87.9	81.7	69.5	67.1	70.3	85.3	84.9	84.3	82.3
DJF	84.7	88.4	87.9	80.6	88.6	80.5	82.4	91.0	89.4	86.8	68.7	45.8	63.2	81.4	89.5	90.4	85.9
MAM	78.1	89.4	83.0	87.3	78.0	54.7	62.7	89.0	84.0	76.0	27.3	39.7	27.1	79.6	64.7	70.2	55.8
JJA	72.9	87.7	78.7	89.5	76.8	91.7	92.8	90.1	90.7	80.5	91.5	88.4	96.3	87.6	91.5	89.7	94.5
SON	75.7	73.8	75.8	87.0	82.5	89.2	95.1	90.5	88.0	82.9	93.0	93.0	95.3	90.6	93.1	88.3	93.0
Cont'd.	1996	1997	1998	1999	200	0 200	1 200	02 200	03 20	04 20	005 20	006	2007	2008	2009	2010	
Cont'd. Jan	1996 88.7	1997 87	1998 82.6									0 06 2	2007 85.1	2008 59.5	2009 52.4	2010 77	
				73.3	87.	8 76	.9 89	0.1 88	3.4 79	9.8 8							
Jan	88.7	87	82.6	73.3	87.	8 76	.9 89	0.1 88 6.9 65	3.4 79 5.6 2 3	9.8 8	7.2 9 8.8	96.9	85.1	59.5	52.4	77	
Jan Feb	88.7 77.9	52.6	82.6 53.7	73.3 48.1 25.3	87. 80. 41.	8 76 4 31 4 0	.9 89 .3 88 .7 48	2.1 88 3.9 65 3.9 25	3.4 79 5.6 2 3	9.8 8 3.3 8 0.0 8	7.2 9 8.8 3.1 8	94	85.1 43.8	59.5 15.9	52.4 15.9	77 23.1	
Jan Feb March	88.7 77.9 65.5	52.6 19.0	82.6 53.7 37.1	73.3 48.1 25.3 17.6	87. 80. 41. 37.	8 76 4 31 4 0 1 26	.9 89 .3 88 .7 48 .2 52	9.1 88 9.9 65 9.9 25 9.6 9	3.4 79 5.6 23 5.6 0.4	9.8 8 3.3 8 0.0 8 0.7 8	7.2 9 8.8 3.1 8 0.4 7	94 94 35.0	85.1 43.8 9.6	59.5 15.9 6.2	52.4 15.9 2.4	77 23.1 13.4	
Jan Feb March April	88.7 77.9 65.5 57.2	87 52.6 19.0 44.4	82.6 53.7 37.1 20.2	73.3 48.1 25.3 17.6 34.9	87. 80. 41. 37. 64.	8 76 4 31 4 0 1 26 7 22	.9 89 .3 88 .7 48 .2 52 .1 82	9.1 88 9.9 65 9.6 9 9.1 82	3.4 79 5.6 23 5.6 0.4 0	9.8 8 3.3 8 0.0 8 0.7 8 3.8 7	7.2 9 8.8 3.1 8 0.4 7 2.9 8	96.9 94 35.0 74.2	85.1 43.8 9.6 34.0	59.5 15.9 6.2 34.1	52.4 15.9 2.4 27.9	77 23.1 13.4 27.7	
Jan Feb March April May	88.7 77.9 65.5 57.2 68.5	87 52.6 19.0 44.4 90.8	82.6 53.7 37.1 20.2 73.2	73.33 48.11 25.33 17.66 34.9	87. 80. 41. 37. 64.	8 76 4 31 4 0 1 26 7 22 4 86	.9 89 .3 88 .7 48 .2 52 .1 82	2.1 88 3.9 65 3.9 25 3.6 9 3.1 82 3.5 92	3.4 79 5.6 23 5.6 0.4 0 62.9 63 62.7 94	9.8 8 3.3 8 0.0 8 0.7 8 3.8 7 4.6 8	7.2 9 8.8 3.1 8 0.4 7 2.9 8 1.0 8	94 94 85.0 74.2 82.8	85.1 43.8 9.6 34.0 47.0	59.5 15.9 6.2 34.1 20.0	52.4 15.9 2.4 27.9 39.0	77 23.1 13.4 27.7 70.5	
Jan Feb March April May June	88.7 77.9 65.5 57.2 68.5 90.2	87 52.6 19.0 44.4 90.8 96.9	82.6 53.7 37.1 20.2 73.2 87.7	73.3 48.1 25.3 17.6 34.9 81.6	87. 80. 41. 37. 64. 83.	8 76 4 31 4 0 1 26 7 22 4 86 4 95	.9 89 .3 88 .7 48 .2 52 .1 82 .6 95 .0 92	2.1 88 3.9 65 3.9 25 3.6 9 3.1 82 3.5 92 3.0 91	3.4 79 5.6 2.5 6.0 2.9 6.2.7 94 1.1 97	9.8 8 3.3 8 9.0 8 9.7 8 9.8 7 9.8 8 9.8 7 9.8 8 9.8 7 9.8 8 9.8 7 9.8 8 9.9 8 9.0 8 9.	7.2 9 8.8 3.1 8 0.4 7 2.9 8 1.0 8 6.9 9	94 94 835.0 74.2 833.1	85.1 43.8 9.6 34.0 47.0 88.6	59.5 15.9 6.2 34.1 20.0	52.4 15.9 2.4 27.9 39.0 66.8	77 23.1 13.4 27.7 70.5 85.7	
Jan Feb March April May June July Aug Sept	88.7 77.9 65.5 57.2 68.5 90.2 95.5	87 52.6 19.0 44.4 90.8 96.9 97.1	82.6 53.7 37.1 20.2 73.2 87.7 88.3	73.3 48.1 25.3 17.6 34.9 81.6 88.4 94.5	87. 80. 41. 37. 64. 83. 91.	8 76 4 31 4 0 1 26 7 22 4 86 4 95 9 94	.9 89 .3 888 .7 48 .2 52 .1 82 .6 95 .0 92 .4 92	8.1 88 8.9 65 8.9 25 8.6 9 8.1 82 8.5 92 8.0 91 8.5 92	3.4 79 5.6 22 5.6 0.4 0.2 6.2.7 94 1.1 95 2.1 95	9.8 8 3.3 8 8 0.0 8 8 7.1 7 5.5 8	7.2 9 8.8 3.1 8 0.4 7 2.9 8 1.0 8 6.9 9 5.1 9	96.9 94 835.0 74.2 32.8 33.1 98.3	85.1 43.8 9.6 34.0 47.0 88.6 91.6	59.5 15.9 6.2 34.1 20.0 66.3 88.3	52.4 15.9 2.4 27.9 39.0 66.8 92.2	77 23.1 13.4 27.7 70.5 85.7 93.5	
Jan Feb March April May June July Aug	88.7 77.9 65.5 57.2 68.5 90.2 95.5 92.1	87 52.6 19.0 44.4 90.8 96.9 97.1 95.9 94.6	82.6 53.7 37.1 20.2 73.2 87.7 88.3 88.8 91.6	73.3 48.1 25.3 17.6 34.9 81.6 88.4 94.5 95.7	87. 80. 41. 37. 64. 83. 91. 93.	8 76 4 31 4 0 1 26 7 22 4 86 4 95 9 94 5 93	.9 89 .3 88 .7 48 .2 52 .1 82 .6 95 .0 92 .4 92 .2 93	88.9 65.9 65.9 25.6 92.1 82.1 82.1 82.1 82.1 82.1 82.1 82.1 8	3.4 79 5.6 22 5.6 0 0.4 0 0.2.9 65 0.1 99 1.1 99 1.1 99 1.1 99 1.1 99	9.8 8 3.3 8 0.0 8 8 0.7 8 7.1 7 7.1 7 7.5.5 8 8.5.1 8	7.2 9 8.8 3.1 8 0.4 7 2.9 8 1.0 8 6.9 9 5.1 9 7.6 8 8.2 9	94	85.1 43.8 9.6 34.0 47.0 88.6 91.6 93.8	59.5 15.9 6.2 34.1 20.0 66.3 88.3 94.5	52.4 15.9 2.4 27.9 39.0 66.8 92.2 90.6	77 23.1 13.4 27.7 70.5 85.7 93.5 89.3	
Jan Feb March April May June July Aug Sept	88.7 77.9 65.5 57.2 68.5 90.2 95.5 92.1 87.6	87 52.6 19.0 44.4 90.8 96.9 97.1 95.9 95.5	82.6 53.7 37.1 20.2 73.2 87.7 88.3 88.8 91.6	73.3 48.1 25.3 17.6 34.9 81.6 88.4 94.5 95.7 94.2	87. 80. 41. 37. 64. 83. 91. 93. 91.	8 76 4 31 4 0 1 26 4 86 4 95 9 94 5 93 3 90	.9 89 .3 88 .7 48 .2 52 .1 82 .6 95 .0 92 .4 92 .2 93 .5 93	2.1 88 3.9 65 3.9 25 3.6 9 2.1 82 3.5 92 3.5 92 3.5 92 3.6 94	3.4 79 3.4 79 3.4 79 3.4 79 3.4 79 3.4 99 4.7 99 4.7 99	9.8 8 3.3 8 0.0 8 0.7 8 4.6 8 7.1 7 5.5 8 5.1 8 5.6 8	7.2 9 8.8 3.1 8 0.4 7 2.9 8 1.0 8 6.9 9 5.1 9 7.6 8 8.2 9	94 85.0 74.2 832.8 833.1 98.3 92.6 88.3	85.1 43.8 9.6 34.0 47.0 88.6 91.6 93.8 92.2	59.5 15.9 6.2 34.1 20.0 66.3 88.3 94.5 91.6	52.4 15.9 2.4 27.9 39.0 66.8 92.2 90.6 90.5	77 23.1 13.4 27.7 70.5 85.7 93.5 89.3 90.9	
Jan Feb March April May June July Aug Sept Oct Nov Dec	88.7 77.9 65.5 57.2 68.5 90.2 95.5 92.1 87.6 91.6 84.5 86.5	87 52.6 19.0 44.4 90.8 96.9 97.1 95.9 94.6	82.6 53.7 37.1 20.2 73.2 87.7 88.3 88.8 91.6	73.3 48.1 25.3 17.6 34.9 81.6 88.4 94.5 95.7 94.2 88.5	87. 80. 41. 37. 64. 83. 91. 93. 91. 95.	8 76 4 31 4 0 1 26 7 22 4 86 4 95 9 94 5 93 3 90 8 86	.9 89 .3 888 .7 48 .2 52 .1 82 .6 95 .0 92 .4 92 .2 93 .5 93 .3 82	2.1 88 3.9 65 3.9 25 3.6 9 2.1 82 3.5 92 3.0 91 3.5 92 3.4 94 4.8 94	3.4 79 2.5.6	9.8 8 8.3.3 8 8.0.0 8 8.0.7 8 8.3.8 7 4.6 8 7.1 7 5.5 8 5.1 8 4.7 9	7.2 9 8.8 3.1 8 0.4 7 2.9 8 1.0 8 6.9 9 5.1 9 7.6 8 8.2 9 6.5 9 0.2	96.9 94 35.0 74.2 32.8 33.1 98.3 92.6 38.3 91.1 93.3 92	85.1 43.8 9.6 34.0 47.0 88.6 91.6 93.8 92.2 91.5	59.5 15.9 6.2 34.1 20.0 66.3 88.3 94.5 91.6 88.0	52.4 15.9 2.4 27.9 39.0 66.8 92.2 90.6 90.5 91.0	77 23.1 13.4 27.7 70.5 85.7 93.5 89.3 90.9 88.6	
Jan Feb March April May June July Aug Sept Oct Nov Dec	88.7 77.9 65.5 57.2 68.5 90.2 95.5 92.1 87.6 91.6 84.5	87 52.6 19.0 44.4 90.8 96.9 97.1 95.9 95.5 94.6	82.6 53.7 37.1 20.2 73.2 87.7 88.3 88.8 91.6 88.5 92.0 92.5	73.3 48.1 25.3 17.6 81.6 88.4 94.5 95.7 94.2 88.5 91.3	87. 80. 41. 37. 64. 83. 91. 93. 91. 93. 95. 76.	8 76 4 31 4 0 1 26 7 22 4 86 4 95 9 94 5 93 3 90 8 86 8 92	.9 89 .3 88 .7 48 .2 52 .1 82 .0 92 .4 92 .2 93 .5 93 .3 82 .3 89	2.1 88 3.9 65 3.9 25 3.6 9 3.6 9 3.1 82 3.5 92 3.0 91 3.5 92 3.4 94 3.8 94 4.7 84 5.8 58	3.4 79 5.6 2.9 6: 2.7 94 1.1 92 1.6 9: 1.7 9: 1.7 9: 1.8 9.9 1.1 94	9.8 8 3.3 8 0.0 8 0.7 8 3.8 7 4.6 8 7.1 7 5.5 8 5.1 8 4.7 9 4.8 9	7.2 9 8.8 3.1 8 0.4 7 2.9 8 1.0 8 6.9 9 5.1 9 7.6 8 8.2 9 6.5 9 0.2	94 94 835.0 74.2 833.1 98.3 92.6 88.3 91.1 93.3	85.1 43.8 9.6 34.0 47.0 88.6 91.6 93.8 92.2 91.5 90.6	59.5 15.9 6.2 34.1 20.0 66.3 88.3 94.5 91.6 88.0 87.3	52.4 15.9 2.4 27.9 39.0 66.8 92.2 90.6 90.5 91.0 88.1	77 23.1 13.4 27.7 70.5 85.7 93.5 89.3 90.9 88.6 88.1	
Jan Feb March April May June July Aug Sept Oct Nov Dec Year DJF	88.7 77.9 65.5 57.2 68.5 90.2 95.5 92.1 87.6 91.6 84.5 86.5 82.2 85.3	87 52.6 19.0 44.4 90.8 96.9 97.1 95.9 95.5 94.6 92.0 84.9	82.6 53.7 37.1 20.2 73.2 87.7 88.3 88.8 91.6 92.0 92.5 74.7	73.3 48.1 25.3 17.6 34.9 81.6 88.4 94.5 95.7 94.2 88.5 91.3 69.5 71.3	87. 80. 41. 37. 64. 83. 91. 93. 91. 95. 76. 78. 86.	8 76 4 31 4 0 1 26 7 22 4 86 4 95 9 94 5 93 3 90 8 86 8 92 0 66	.9 89 .3 88 .7 48 .2 52 .1 82 .6 95 .0 92 .4 92 .2 93 .5 93 .3 82 .3 89 .3 83	2.1 88 3.9 65 3.9 25 3.6 9 3.6 9 3.5 92 3.0 91 3.5 92 3.4 94 3.7 84 3.8 58 3.4 73 3.1 81	3.4 79 5.6 2.9 6.5 1.1 92 1.1 92 1.1 94 1.1 94 1.1 94 1.1 94 1.1 94 1.1 94 1.1 94 1.1 94 1.1 94 1.1 94 1.1 94	9.8 8 3.3 8 0.0 8 0.7 8 3.8 7 4.6 8 7.1 7 5.5 8 5.1 8 4.7 9 4.8 9 9.6 8 3.7 9	7.2 9 8.8 3.1 8 0.4 7 2.9 8 1.0 8 6.9 9 5.1 9 7.6 8 8.2 9 6.5 9 0.2 4.8 8 0.3 9	94 94 94 94 94 94 94 94	85.1 43.8 9.6 34.0 47.0 88.6 91.6 93.8 92.2 91.5 90.6 95.1 71.9 73.6	59.5 15.9 6.2 34.1 20.0 66.3 88.3 94.5 91.6 88.0 87.3 89.7 61.8 56.8	52.4 15.9 2.4 27.9 39.0 66.8 92.2 90.6 90.5 91.0 88.1 90.1	77 23.1 13.4 27.7 70.5 85.7 93.5 89.3 90.9 88.6 88.1 73.6 68.5 63.4	
Jan Feb March April May June July Aug Sept Oct Nov Dec Year DJF MAM	88.7 77.9 65.5 57.2 68.5 90.2 95.5 92.1 87.6 91.6 84.5 86.5	87 52.6 19.0 44.4 90.8 96.9 97.1 95.9 95.5 94.6 92.0 84.9 79.2 75.4 51.4	82.6 53.7 37.1 20.2 87.7 88.3 88.8 91.6 88.5 92.0 92.5 74.7 73.7 43.5	73.3 48.1 25.3 17.6 34.9 81.6 88.4 94.5 95.7 94.2 88.5 91.3 69.5 71.3	87. 80. 41. 37. 64. 83. 91. 93. 91. 95. 76. 78. 86.	8 76 4 31 4 0 1 26 7 22 4 86 4 95 9 94 5 93 3 90 8 86 8 92 0 66 5 61	.9 89 .3 88 .7 48 .2 52 .1 82 .6 95 .0 92 .4 92 .2 93 .5 93 .3 82 .3 89 .3 89	2.1 88 3.9 65 3.9 25 3.6 9 3.6 9 3.5 92 3.0 91 3.5 92 3.4 94 3.7 84 3.8 58 3.4 73 3.1 81	3.4 79 5.6 2.9 65 6.1 99 1.1 99 1.1 99 1.1 94 1.1 94 1.1 94 1.1 94 1.1 94 1.1 94 1.1 94 1.1 94 1.1 95 1.	9.8 8 3.3 8 9.0 8 9.7 8 3.8 7 4.6 8 7.1 7 5.5 8 5.1 8 5.6 8 4.7 9 9.6 8 3.7 9	7.2 9 8.8 3.1 8 0.4 7 2.9 8 1.0 8 6.9 9 5.1 9 7.6 8 8.2 9 6.5 9 0.2 4.8 8 8.8 8	94 94 94 94 94 94 94 94	85.1 43.8 9.6 34.0 47.0 88.6 91.6 93.8 92.2 91.5 90.6 95.1 71.9	59.5 15.9 6.2 34.1 20.0 66.3 88.3 94.5 91.6 88.0 87.3 89.7 61.8	52.4 15.9 2.4 27.9 39.0 66.8 92.2 90.6 90.5 91.0 88.1 90.1 62.2	77 23.1 13.4 27.7 70.5 85.7 93.5 89.3 90.9 88.6 88.1 73.6 68.5	
Jan Feb March April May June July Aug Sept Oct Nov Dec Year DJF	88.7 77.9 65.5 57.2 68.5 90.2 95.5 92.1 87.6 91.6 84.5 86.5 82.2 85.3	87 52.6 19.0 44.4 90.8 96.9 97.1 95.9 95.5 94.6 92.0 84.9 79.2 75.4	82.6 53.7 37.1 20.2 87.7 88.3 88.8 91.6 88.5 92.0 92.5 74.7 73.7 43.5 88.3	73.3 48.1 25.3 17.6 34.9 81.6 88.4 94.5 95.7 94.2 88.5 91.3 69.5 71.3 25.9 88.2	87. 80. 41. 37. 64. 83. 91. 93. 95. 76. 78. 86.	8 76 4 31 4 0 1 26 7 22 4 86 4 95 9 94 5 93 3 90 8 86 8 92 0 66 5 61 7 16	.9 89 .3 88 .7 48 .2 52 .1 82 .6 95 .0 92 .4 92 .2 93 .5 93 .3 82 .3 89 .3 61	2.1 88 3.9 65 3.9 25 3.6 9 2.1 82 3.5 92 3.0 91 3.5 92 3.4 94 4.8 94 4.7 84 9.8 58 3.4 73 3.5 92 3.1 82 3.5 92 3.1 82 3.5 92 3.1 82 3.1 82 3.1 82 3.1 82 3.2 82 3.3 82 3.4 94 3.5 82 3.6 83 3.7 84 3.8	3.4 79 5.6 2.7 9.4 (0.2.7 9.4.1 9.3.1 9.4.1 9.3.3 69 3.3 69 3.3 69 3.3 69 3.3 69	9.8 8 3.3 8 0.0 8 0.7 8 4.6 8 7.1 7 5.5 8 5.6 8 4.7 9 4.8 9 9.6 8 9.6 8 7.1 7	7.2 9 8.8 3.1 8 0.4 7 2.9 8 1.0 8 6.9 9 5.1 9 7.6 8 8.2 9 6.5 9 0.2 4.8 8 8.8 8	94 94 94 94 94 94 94 94	85.1 43.8 9.6 34.0 47.0 88.6 91.6 93.8 92.2 91.5 90.6 95.1 71.9 73.6	59.5 15.9 6.2 34.1 20.0 66.3 88.3 94.5 91.6 88.0 87.3 89.7 61.8 56.8	52.4 15.9 2.4 27.9 39.0 66.8 92.2 90.6 90.5 91.0 88.1 90.1 62.2 52.7	77 23.1 13.4 27.7 70.5 85.7 93.5 89.3 90.9 88.6 88.1 73.6 68.5 63.4	

(Box W in Fig. 3.) Mean ice concentrations of less than 50% are in boldface and boxed to show the increase in frequency of occurrences of low sea ice concentration in the later (post-1989 and particularly post-1997) epoch versus the earlier (pre-1989) epoch. The seasonal data are plotted in Fig. 4d.



An Early Cretaceous eutherian and the placental-marsupial dichotomy

Shundong Bi^{1,2,3}*, Xiaoting Zheng^{4,5}, Xiaoli Wang^{4,5}*, Natalie E. Cignetti², Shiling Yang^{6,7} & John R. Wible³*

Molecular estimates of the divergence of placental and marsupial mammals and their broader clades (Eutheria and Metatheria, respectively) fall primarily in the Jurassic period. Supporting these estimates, Juramaia—the oldest purported eutherian—is from the early Late Jurassic (160 million years ago) of northeastern China. Sinodelphys—the oldest purported metatherian—is from the same geographic area but is 35 million years younger, from the Jehol biota. Here we report a new Jehol eutherian, Ambolestes zhoui, with a nearly complete skeleton that preserves anatomical details that are unknown from contemporaneous mammals, including the ectotympanic and hyoid apparatus. This new fossil demonstrates that Sinodelphys is a eutherian, and that postcranial differences between Sinodelphys and the Jehol eutherian Eomaia—previously thought to indicate separate invasions of a scansorial niche by eutherians and metatherians—are instead variations among the early members of the placental lineage. The oldest known metatherians are now not from eastern Asia but are 110 million years old from western North America, which produces a 50-million-year ghost lineage for Metatheria.

Over the past 20 years, more than 120 new genera of vertebrates have been described from the Early Cretaceous Jehol biota of northeastern China¹. However, therian mammals—which include living placentals and marsupials as well as their respective broader clades, Eutheria and Metatheria—are rare at Jehol, and are represented by only three monotypic genera. Two of these genera, *Eomaia*² and *Acristatherium*³, are widely accepted as eutherians^{2–5}, although the former was placed outside of Theria in a recent study⁶ that was limited taxonomically by its inclusion of only three of the more than 40 known Cretaceous eutherians. The third Jehol therian, *Sinodelphys*⁷, is generally considered to be the oldest known metatherian and to therefore provide morphological evidence for divergence of these two clades of living Theria. Here we describe a fourth Jehol therian and incorporate it into a phylogenetic analysis that supports all four Jehol therians as stem placentals.

Class Mammalia Linnaeus, 1758 Infraclass Eutheria sensu Huxley, 1880 Order incertae sedis Family incertae sedis Ambolestes gen. nov. Ambolestes zhoui sp. nov.

Etymology. *Ambo* (Latin), both, in reference to the mixture of features previously held⁷ to be from eutherians and metatherians; *lestes* (Greek), robber, a common ending for Cretaceous eutherians. The specific name *zhoui* is given in reference to Zhonghe Zhou, for his pioneering studies of the Jehol biota.

Holotype. A nearly complete skeleton preserved on main and counterpart slabs A and B of STM33-5 (Tianyu Museum of Nature, Linyi, Shandong Province, China) (Fig. 1, Extended Data Fig. 1 and Supplementary Tables 1, 2).

Locality and horizon. The holotype is from the Lower Cretaceous Yixian Formation at Xisanjia, Inner Mongolia, China, dated to about 126 million years ago (Ma)⁸.

Diagnosis. Tooth formula: I?:C1:P5:M3/I2:C1:P4:M3 (I, incisor; C, canine; P, premolar; M, molar; superscript and subscript denote upper and lower teeth, respectively) (Fig. 2 and Extended Data Figs. 2, 3). Differs from metatherians in that it has eight upper postcanine teeth, whereas metatherians have seven^{9,10}. Differs from Jurassic and other Early Cretaceous eutherians known from full lower postcanine dentitions^{2-5,11}—except for Sinodelphys (Extended Data Figs. 4, 5 and Supplementary Information) and Sasayamamylos¹²—in that it has seven lower postcanine teeth, whereas Jurassic and other Early Cretaceous eutherians have eight. Differs from Sinodelphys⁷, Sasayamamylos¹² and Montanalestes¹³ that possess a slightly medially inflected mandibular angle and metatherians^{9,10} that possess a fully medially inflected angle: Ambolestes has a posteriorly directed mandibular angle as in Eomaia² and Prokennalestes¹¹. Differs from Sinodelphys⁷ in that it has a large trapezium in the wrist (Fig. 3a). Differs from Sinodelphys⁷ and Eomaia² in that it has double-rooted upper canines. Differs from Sinodelphys⁷, Eomaia² and Acristatherium³ in that it has a first upper premolar that is smaller than the second. Differs from Sinodelphys⁷, Juramaia⁵ and *Eomaia*² in that it has a less-developed supraspinous fossa of the scapula and an enlarged parafibula. Differs from Juramaia⁵, Prokennalestes and Late Cretaceous eutherians^{4,11} in that it lacks a protocone or protoconal swelling on the ultimate upper premolar. Differs from Sasayamamylos¹² in that it has a lingual cingulid on the ultimate lower premolar, a Meckelian sulcus and masseteric foramen on the dentary (Extended Data Fig. 2). Differs from *Durlstotherium* and *Durlstodon* in that it has upper molars with a lower protocone and no conules¹⁴. Differs from the non-therians Aegialodon 15, Kielantherium, Vincelestes, Peramus and Nanolestes^{4,11} in that it has a more-closed trigonid configuration on the lower molars.

Phylogenetic analyses

Phylogenetic analysis using parsimony (Supplementary Information) places the four Jehol therians (*Ambolestes*, *Sinodelphys*, *Acristatherium* and *Eomaia*) within Eutheria (Fig. 4 and Extended Data Fig. 6). The

¹Yunnan Key Laboratory for Palaeobiology, Yunnan University, Kunming, China. ²Department of Biology, Indiana University of Pennsylvania, Indiana, PA, USA. ³Carnegie Museum of Natural History, Pittsburgh, PA, USA. ⁴Institute of Geology and Paleontology, Linyi University, Linyi, China. ⁵Tianyu Museum of Nature, Linyi, China. ⁶Key Laboratory of Cenozoic Geology and Environment, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China. ⁷CAS Center for Excellence in Life and Paleoenvironment, Beijing, China. *e-mail: shundong.bi@iup.edu; wang_7355@163.com; wiblej@carnegiemnh.org

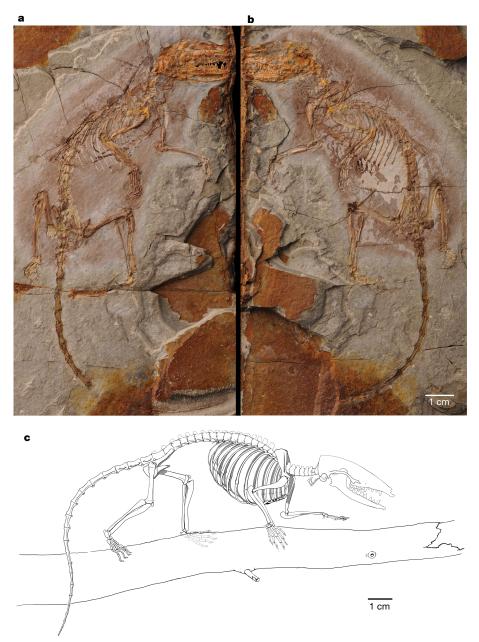


Fig. 1 | Holotype specimen of A. zhoui STM33-5, Tianyu Museum of Nature, Shandong Province, China. a, Main slab A. b, Counterpart slab B. c. Restoration.

first three are in a clade along with Montanalestes from the Early Cretaceous period of western North America¹³, with Ambolestes in a sister relationship to Sinodelphys. The fourth Jehol therian, Eomaia, and the 160-million-year-old Juramaia—from the Tiaojishan Formation of northeastern China⁵—are sister to the clade that includes Placentalia. The major departure in our analyses that include Ambolestes from previous studies with a similar phylogenetic scope is the position of Sinodelphys, which had previously been identified at the base of Metatheria^{5,7} and used to support an Asian origin for this clade⁷. Moreover, features of the wrist and ankle that distinguished Sinodelphys from the roughly contemporaneous Eomaia were used to support a fundamental dichotomy in the ways in which metatherians and eutherians attained their scansorial locomotory modes⁷. With Sinodelphys moved to Eutheria in our analysis, the oldest metatherians are represented by the sub-clades Deltatheroida and Marsupialiformes, which are both from the Albian (approximately 110 Ma) of western North America 16,17 (Fig. 4); the earliest members of these sub-clades from Asia are at least 20 million years younger^{18,19}. Additionally, in the context of the new phylogenetic arrangement, the postcranial differences

between *Sinodelphys* and *Eomaia* are not the result of metatherian-eutherian cladogenesis as had previously been suggested⁷, but are instead variations within early eutherians. In the wrist example (Fig. 3a), *Sinodelphys* does share a large scaphoid, triquetrum, and hamate and small trapezium with many extant marsupials, and *Eomaia* shares the opposite state for these bones with some extant placentals. *Ambolestes* has a mixture of features of the two; it resembles *Sinodelphys* in possessing a large scaphoid, triquetrum and hamate, and *Eomaia* in possessing a large trapezium. The earliest members of Deltatheroida and Marsupialiformes are known only from isolated teeth and fragmentary jaws^{16,17}, and therefore provide no evidence of early metatherian locomotory adaptations.

Morphology of Ambolestes

In extant mammals, the ectotympanic bone is a component of the sound-conducting apparatus of the middle ear that provides an attachment for the tympanic membrane and the malleus—the latter either directly through bone (Fig. 5), or indirectly by ligament. Because of its delicate nature and because it is usually only loosely attached to

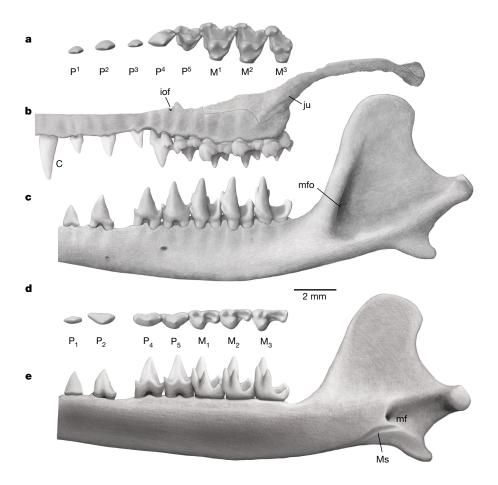


Fig. 2 | Dental and upper and lower jaw features of A. zhoui STM33-5. a, Left upper postcanine dentition in occlusal view. b, Left upper jaw in lateral view. c, Left dentary in lateral view. **d**, Left lower postcanine dentition in occlusal view. e, Right dentary in medial view. Lateral views in b, c are based on direct observation of the specimen; occlusal and lingual views in a, d, e are reconstructions based on direct observation of the specimen and on computed tomography scans. C, upper canine; iof, infraorbital foramen; ju, jugal; mf, mandibular foramen; mfo masseteric foramen; Ms, Meckel's sulcus. For the postcanine teeth: M, molar; P, premolar; and superscript and subscript numbers denote upper and lower teeth, respectively.

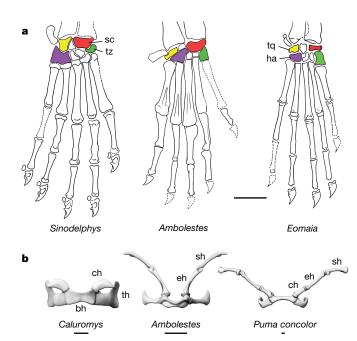


Fig. 3 | Forefoot and hyoid apparatus of *A. zhoui* compared with those of other mammals. a, Forefoot in dorsal view, of *Sinodelphys* (redrawn from a previous study⁷), *Ambolestes* STM33-5 and *Eomaia* (redrawn from a previous study⁷). b, Hyoid apparatus in ventral view in the extant didelphid marsupial *Caluromys derbianus* (Carnegie Museum of Natural History CM 119051), *Ambolestes* STM33-5 and the extant carnivoran *Puma concolor* (Carnegie Museum of Natural History CM 59940). bh, basihyal; ch, ceratohyal; eh, epihyal; ha, hamate; sc, scaphoid; sh, stylohyal; th, thyrohyal; tq, triquetrum; tz, trapezium. Scale bars, 2 mm.

other bones, the ectotympanic is poorly known in Mesozoic mammaliaforms and our understanding of the evolution of this bone is therefore extremely limited. Ambolestes preserves both the left (Extended Data Fig. 2) and right ectotympanic bones. This C-shaped bone has a simple ring-shaped anterior crus (leg) and an expanded posterior crus (Fig. 5); the reverse condition occurs in the Late Cretaceous eutherians Asioryctes²⁰, Uchkudukodon²¹ and Zalambdalestes²², and in most extant therians. In the proportions of its crura, the ectotympanic of Ambolestes closely resembles that of some extant didelphid marsupials (opossums), especially Monodelphis (Fig. 5). As in opossums, the ectotympanic of Ambolestes has a distinct facet that curves around the rostral surface of the anterior crus (also seen in *Uchkudukodon*²¹), which held a similarly curved anterior process of the malleus (Fig. 5). In the C-shaped ectotympanic and curved anterior process of the malleus, Ambolestes and Monodelphis differ from non-therians such as the extant monotreme Tachyglossus and the eutriconodont Liaoconodon²³, in which the ectotympanic shape is not as uniformly curved and the anterior process of the malleus is straight (Fig. 5). These characteristics of the ectotympanic and malleus in Tachyglossus and Liaoconodon are retentions of the primitive arrangement of homologous structures in Mesozoic mammaliaforms such as Morganucodon, in which there was an osseous connection to the lower jaw through the postdentary bones²⁴. This connection is retained in *Liaoconodon* through the ossified Meckel's cartilage²³ but is lost in Tachyglossus, as it is in all extant mammals. The short, shallow Meckelian sulcus in Ambolestes (Fig. 2) is a primitive retention that is lost in most therians, and its ectotympanic was probably fully suspended from the skull base as in extant mammals. The phylogenetic transformation of the shape of the ectotympanic and anterior process of the malleus from the straight non-therian to the curved therian pattern is traceable in the ontogeny of extant marsupials²⁵ (Fig. 5).

The hyoid apparatus in extant mammals sits in the floor of the mouth and provides attachment for muscles of the tongue, pharynx and body

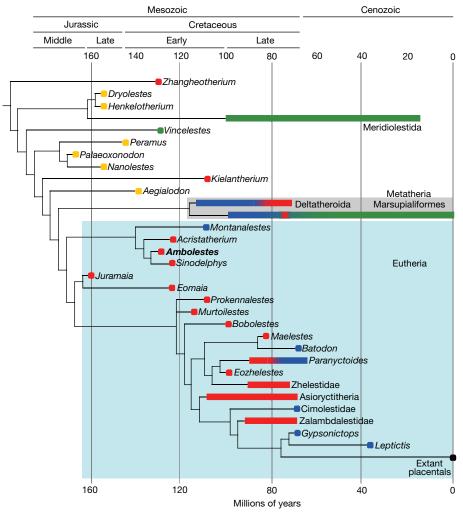


Fig. 4 | **Relationships of** *A. zhoui* **to other mammals.** Simplified tree of the strict consensus of the four most-parsimonious trees (Supplementary Information). Colours indicate current continent of discovery for the specimens studied here (except Placentalia): red, Asia; yellow,

Europe; green, South America; blue, North America. Deltatheroida, Marsupialiformes and *Paranyctoides* are mixed, with specimens found on different continents.

wall²⁶. Three bones (the unpaired basihyal and paired thyrohyals) form a basal series possessed by all mammals. Four paired bones—the ceratohyals, epihyals, stylohyals and tympanohyals (for example, Puma in Fig. 3b)—are variably present and may form a suspensory series; tympanohyals are usually fused to the petrosal bone on the skull base. In situations in which one or more parts of the suspensory series fail to ossify (for example, Caluromys in Fig. 3b), the stylohyoid ligament connects the basal series to the basicranium. The number of individual bones varies from nine (for example, Puma in Fig. 3b) to as few as one (as in humans, albeit one formed from six ossification centres²⁷). The number and shape of the elements has some phylogenetic value as, for example, all marsupials have the same five elements²⁶ (see *Caluromys* in Fig. 3b). The evolution of the mammalian hyoid is poorly known because few of these bones are preserved in the fossil record. Possible hyoid fragments have been noted for some Mesozoic mammals (including Maotherium²⁸, Liaoconodon²³ and Vilevolodon²⁹) but a complete series has not previously been reported. The hyoid apparatus in Ambolestes is complete and composed of seven elements (Fig. 3b). Of the extant mammals studied to date, the only ones to possess the same elements as those present in Ambolestes are in the squirrel family, including all tree squirrels and some terrestrial squirrels³⁰. In the squirrels, the unpaired basihyal and paired thyrohyals, epihyals and stylohyals³⁰ are present, and we identify the elements in Ambolestes accordingly. As with other parts of the anatomy, the hyoid of Ambolestes has a mosaic of marsupial and placental features. The basal series of *Ambolestes* resembles that of marsupials in terms of its basihyal, which is much shorter than the

thyrohyals, but differs from marsupials and resembles many placentals in having a continuous ossified suspensory series (Fig. 3b).

With an estimated body weight of 34-44 g (Supplementary Information)—roughly equivalent to an extant mouse opossum (Marmosa)—Ambolestes is similar in size to other Jehol eutherians^{2,3,7}, whereas the Jurassic *Juramaia* is smaller at 15–17 g⁵. *Eomaia* and Sinodelphys, both of which are known from postcranial bones, were originally suggested to be scansorial, primarily on the basis of their manual and pedal digit proportions^{2,7}. A recent and more comprehensive study of these elements identified Eomaia as arboreal and Sinodelphys as scansorial³¹. The manual digit proportions and intermembral index of Ambolestes resemble those of Eomaia more than those of Sinodelphys (Extended Data Fig. 7); therefore, depending on which authors are followed, Ambolestes may be regarded as either scansorial or arboreal—categories that do not in any case separate cleanly with most metrics³¹. Regardless of what may have been happening elsewhere, the taxa from the Jehol biota demonstrate that arboreality was an important theme in the diversification of early Eutheria.

Eutherian-metatherian divergence

Merging recent molecular estimates for the divergence of Eutheria and Metatheria^{32,33} results in a broad temporal range for this event, from the earliest Cretaceous to the latest Triassic period (140–215 Ma). Only three therian fossils, each purported to be a eutherian^{5,14}, have been found within this range and therefore have the potential to provide some support for the molecular estimates.

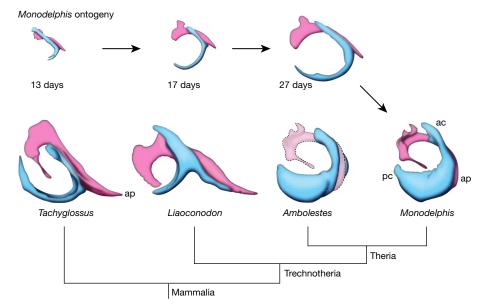


Fig. 5 | Ectotympanic and malleus in $A.\ zhoui$ and other mammals. Bottom row shows ectotympanic (blue) and malleus (pink) of the extant monotreme $Tachyglossus\ aculeatus$ (Carnegie Museum of Natural History CM 50809), the Jehol eutriconodont $Liaoconodon\ hui$ (redrawn from a previous study²³), Ambolestes, and the extant didelphid $Monodelphis\ domestica$, (Carnegie Museum of Natural History CM 80017). Not to scale.

with adult *Monodelphis* in bottom row. The malleus is not known for *Ambolestes*: the anterior process is reconstructed based on the facet on the anterior crus of the ectotympanic. ac, anterior crus of ectotympanic; ap, anterior process of malleus; pc, posterior crus of ectotympanic.

Top row shows *M. domestica* at postnatal days 13, 17 and 27 (from a previously published computed tomography dataset²⁵). Shown to scale

These are 160-million-year-old *Juramaia*, represented by a relatively complete dentition⁵ (Extended Data Fig. 4), and the 145-millionyear-old Durlstotherium and Durlstodon from southern England, each of which is represented by only an incomplete upper ultimate molar¹⁴. Despite the 35 million years that separate *Juramaia* from the Jehol eutherian *Eomaia*, their dentitions share numerous similarities that form the primary basis for their sister-group relationship in our analysis (Fig. 4). Durlstotherium and Durlstodon exhibit a similar phenomenon; their upper ultimate molars are unexpectedly advanced and resemble those of younger taxa. However, for *Durlstotherium* and Durlstodon this separation is more in the range of 50 million years, as the younger taxa are early Late Cretaceous eutherians with upper ultimate molars that have conules and high protocones^{4,11}. How can these unexpected resemblances across vast geological time be explained? For Juramaia, the stratigraphic position of the only known specimen has been questioned³⁴, with a suggested younger age that appears to have been based on the notable similarities to geologically younger taxa. An alternative explanation for these unexpected resemblances is the early appearance of a derived morphology coupled to a slow rate of dental change. However, with so few data points across this initial phase of eutherian evolutionary history an informed conclusion is not possible. If the molecular estimates are appropriate, mammalian palaeontologists face the challenge of finding more fossils that are contemporaneous with the molecular estimates for the eutherian side—and finding any fossils for the metatherian side, for which a 50-million-year ghost lineage now exists.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0210-3.

Received: 27 April 2018; Accepted: 10 May 2018; Published online: 13 June 2018

- Zhou, Z. & Wang, Y. Vertebrate diversity of the Jehol Biota as compared with other lagerstätten. Sci. China Earth Sci. 53, 1894–1907 (2010).
- 2. Ji, Q. et al. The earliest known eutherian mammal. *Nature* **416**, 816–822 (2002).

- Hu, Y., Meng, J., Li, C. & Wang, Y. New basal eutherian mammal from the Early Cretaceous Jehol biota, Liaoning, China. Proc. R. Soc. Lond. B 277, 229–236 (2010).
- Wible, J. R., Rougier, G. W., Novacek, M. J. & Asher, R. J. The eutherian mammal *Maelestes gobiensis* from the Late Cretaceous of Mongolia and the phylogeny of Cretaceous Eutheria. *Bull. Am. Mus. Nat. Hist.* 327, 1–123 (2009).
- Luo, Z.-X., Yuan, C.-X., Meng, Q.-J. & Ji, Q. A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature* 476, 442–445 (2011).
- O'Leary, M. A. et al. The placental mammal ancestor and the post-K-Pg radiation of placentals. Science 339, 662–667 (2013).
- Luo, Z.-X., Ji, Q., Wible, J. R. & Yuan, C.-X. An Early Cretaceous tribosphenic mammal and metatherian evolution. Science 302, 1934–1940 (2003).
- 8. Chang, S.-C., Gao, K.-Q., Zhou, C.-F. & Jourdan, F. New chronostratigraphic constraints on the Yixian Formation with implications for the Jehol Biota. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **487**, 399–406 (2017).
- Williamson, T. E., Brusatte, S. L. & Wilson, G. P. The origin and early evolution of metatherian mammals: the Cretaceous record. ZooKeys 465, 1–76 (2014).
- Bi, S., Jin, X., Li, S. & Du, T. A new Cretaceous metatherian mammal from Henan, China. PeerJ 3, e896 (2015).
- Averianov, A. O. & Archibald, D. in Legacy of the Gobi Desert: Papers in Memory of Zofia Kielan-Jaworowska (Palaeontologia Polonica 67) (eds Cifelli, R. L. & Fostowicz-Frelik, Ł.) 25–33 (Institute of Paleobiology of the Polish Academy of Sciences. Warsaw. 2016).
- Kusuhashi, N. et al. A new Early Cretaceous eutherian mammal from the Sasayama Group, Hyogo, Japan. Proc. R. Soc. Lond. B 280, 20130142 (2013).
- Cifelli, R. L. Tribosphenic mammal from the North American Early Cretaceous. Nature 401, 363–366 (1999).
- Sweetman, S. C., Smith, G. & Martill, D. M. Highly derived eutherian mammals from the earliest Cretaceous of southern Britain. Acta Palaeontol. Pol. 62, 657–665 (2017).
- Kermack, K. A., Lees, P. M. & Mussett, F. Aegialodon dawsoni, a new trituberculosectorial tooth from the Lower Wealden. Proc. R. Soc. Lond. B 162, 535–554 (1965).
- Davis, B. M., Cifelli, R. L. & Kielan-Jaworowska, Z. in Mammalian Evolutionary Morphology: a Tribute to Frederick S. Szalay (eds Sargis, E. J. & Dagosto, M.) 3–24 (Springer, Dordrecht, 2008).
- Čifelli, R. L. & Davis, B. M. Tribosphenic mammals from the Lower Cretaceous Cloverly Formation of Montana and Wyoming. *J. Vertebr. Paleontol.* 35, e920848 (2015).
- Áverianov, A. O., Archibald, J. D. & Ekdale, E. G. New material of the Late Cretaceous deltatheroidan mammal Sulestes from Uzbekistan and phylogenetic reassessment of the metatherian–eutherian dichotomy. J. Syst. Palaeontol. 8, 301–330 (2010).
- Szalay, F. S. & Trofimov, B. A. The Mongolian Late Cretaceous Asiatherium, and the early phylogeny and paleobiogeography of Metatheria. J. Vertebr. Paleontol. 16, 474–509 (1996).
- Kielan-Jaworowska, Z. in Results of the Polish-Mongolian Palaeontological Expeditions. Part IX. (Palaeontologia Polonica 42) (ed. Kielan-Jaworowska, Z.) 25–78 (Institute of Paleobiology of the Polish Academy of Sciences, Warsaw, 1981).

RESEARCH ARTICLE

- 21. McKenna, M. C., Kielan-Jaworowska, Z. & Meng, J. Earliest eutherian mammal skull from the Late Cretaceous (Coniacian) of Uzbekistan. Acta Palaeontol. Pol. **45**, 1-54 (2000).
- Wible, J. R., Novacek, M. J. & Rougier, G. W. New data on the skull and dentition in the Mongolian Late Cretaceous eutherian mammal Zalambdalestes. Bull. Am. Mus. Nat. Hist. **281**, 1–144 (2004).
- 23. Meng, J., Wang, Y. & Li, C. Transitional mammalian middle ear from a new Cretaceous Jehol eutriconodont. Nature 472, 181–185 (2011).
- 24. Luo, Z.-X. Developmental patterns in Mesozoic evolution of mammal ears. Annu. Rev. Ecol. Evol. Syst. 42, 355–380 (2011).
- 25. Ramírez-Chaves, H. E. et al. Mammalian development does not recapitulate suspected key transformations in the evolutionary detachment of the mammalian middle ear. Proc. R. Soc. Lond. B 283, 20152606 (2016).
- 26. Gasc, J.-P. in Traité de Zoologie tome XVI, fasc. 1 (ed. Grassé, P.-P.) 550-583, 1103-1106 (Masson, Paris, 1971).
- Standring, S. (ed.) Gray's Anatomy: the Anatomical Basis of Clinical Practice 40th edn (Churchill Livingstone, Edinburgh, 2008).
- 28. Rougier, G. W., Ji, Q. & Novacek, M. J. A new symmetrodont mammal with fur
- impressions from the Mesozoic of China. Acta Geol. Sin. 77, 7-14 (2003). 29. Luo, Z.-X. et al. New evidence for mammaliaform ear evolution and feeding adaptation in a Jurassic ecosystem. Nature 548, 326-329 (2017).
- Hoffmeister, R. G. & Hoffmeister, D. F. The hyoid in North American squirrels, Sciuridae, with remarks on associated musculature, An. Inst. Biol. Univ. Nac. Auton. Mex. Ser. Zool. **62**, 219–234 (1991).
- 31. Chen, M. & Wilson, G. P. A multivariate approach to infer locomotor modes in Mesozoic mammals. *Paleobiology* **41**, 280–312 (2015).
- 32. Meredith, R. W. et al. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334, 521-524 (2011).
- 33. dos Reis, M. et al. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc. R. Soc. Lond. B 279, 3491-3500 (2012).
- 34. Meng, J. Mesozoic mammals of China: implications for phylogeny and early evolution of mammals. Natl Sci. Rev. 1, 521-542 (2014).

Acknowledgements We thank S. Xie for specimen preparation; P. Bowden for illustration; W. Gao for photography; Y. Hou and P. Yin for computed tomography scanning; D. Koyabu and V. Weisbecker for providing the computed tomography dataset of Monodelphis; and J. Meng, X. Xu, B. Jiang and Y. Huang for assistance and discussion. The study was supported by the National Natural Science Foundation of China (41688103, 41728003, 41372014, 41472023) and Chinese Academy of Sciences (XDPB0503). Support for S.B. was provided by the MEC International Joint Laboratory for Palaeobiology and Palaeoenvironment, Yunnan University. Support for J.R.W. is provided by the National Science Foundation Grant DEB 1654949 and Carnegie Museum of Natural History.

Reviewer information Nature thanks R. Cifelli, D. Krause and G. Rougier for their contribution to the peer review of this work.

Author contributions S.B. and J.R.W. conceived the study, undertook comparative and analytical work and wrote the paper; N.E.C. performed the ternary plot; and X.Z., S.Y. and X.W. contributed to fossil interpretation and provided feedback to the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0210-3

Supplementary information is available for this paper at https://doi. org/10.1038/s41586-018-0210-3.

Reprints and permissions information is available at http://www.nature.com/ reprints.

Correspondence and requests for materials should be addressed to S.B. or X.W. or J.R.W.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Computed tomography scanning. The skull of the main slab (A) of *Ambolestes* STM33-5 (Fig. 1a and Extended Data Fig. 1a) was scanned with the 225-kV micro-CT at the Key Laboratory of Vertebrate Evolution and Human Origin of CAS. A total of 1,797 transmission images were reconstructed in a 518 \times 518 matrix of 727 slices in two-dimensional reconstruction software developed by the Institute of High Energy Physics, CAS. The three-dimensional reconstructions were created with the software Mimics (version 16.1).

Phylogenetic analysis. The data matrix consisting of 64 taxa and 401 characters was analysed in TNT 35 under the new technology search (sectorial search, ratchet, tree fusing) set to 100 iterations, followed by a traditional search. All characters were equally weighted and non-additive. The search procedure resulted in four most-parsimonious trees (MPTs) of length 1,776 (consistency index = 0.319; retention index = 0.598). The strict consensus tree (1,779 steps; consistency index = 0.318; retention index = 0.597) of these four MPTs is presented in Fig. 4. The list of the synapomorphies common to the four MPTs was produced using

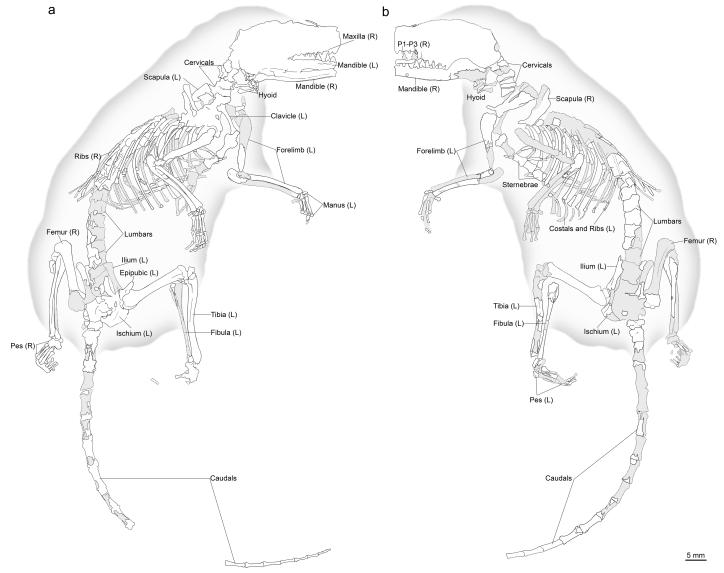
the 'Common Synapomorphies' options in TNT. A list of synapomorphies for the clades Eutheria, Metatheria and the clade formed by Ambolestes and Sinodelphys is provided in the Supplementary Information.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

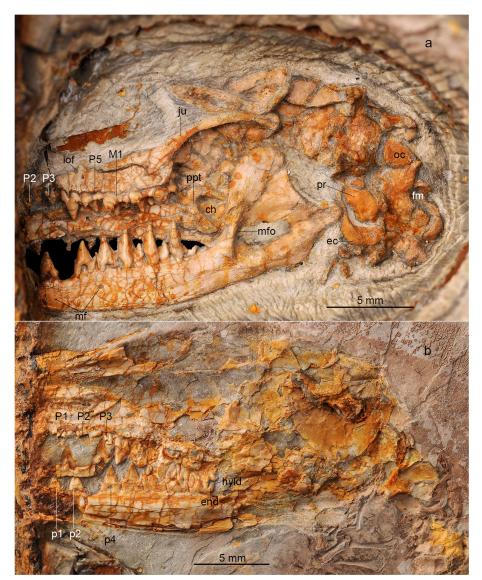
Data availability. The new specimens analysed in this study have been deposited at the Tianyu Museum of Nature, Linyi, Shandong Province, China. Graphics and phylogenetics data are provided in the Supplementary Information. Life Science Identifiers (LSID) for the new genus and species are registered with Zoobank (http://zoobank.org): urn:lsid:zoobank.org:act:6DC5CD12-44DE-4C5E-B90C-EDA0481695E3 and urn:lsid:zoobank.org:act:4A18FEDF-039B-46F7-836F-A6B147E50DCC. The data matrix for the phylogenetic analysis has been deposited in MorphoBank (http://morphobank.org/permalink/?P2799).

35. Goloboff, P. A., Farris, J. S. & Nixon, K. C. TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–786 (2008).



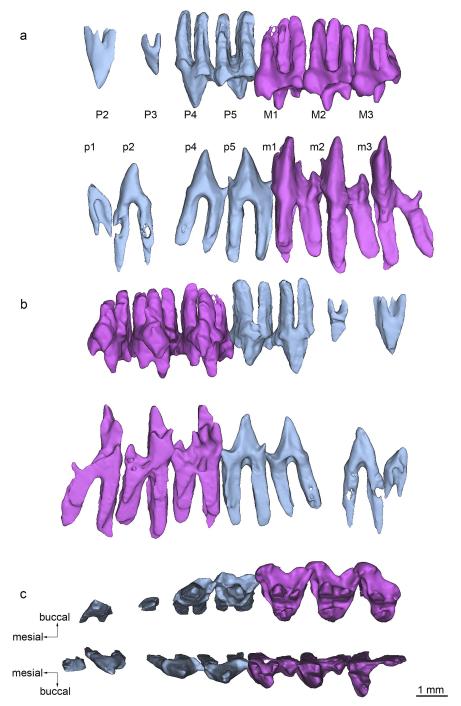


Extended Data Fig. 1 | Line drawings of A. zhoui STM33-5. a, Main slab. b, Counterpart slab.



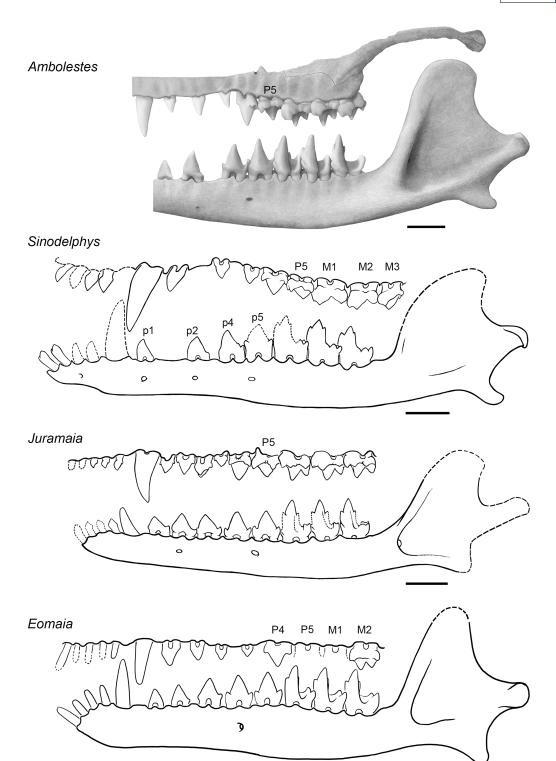
Extended Data Fig. 2 | Close-up views of craniodental features of *A. zhoui* STM33-5. a, Partial skull and left dentary of the main slab in lateral view. b, Partial skull and right dentary of the counterpart slab in medial view. M and m, upper and lower molars, respectively; P and p,

upper and lower premolars, respectively. ch, choanae; ec, ectotympanic; end, entoconid; fm, foramen magnum; hyld, hypoconulid; iof, infraorbital foramen; ju, jugal; mf, mental foramina; mfo, masseteric foramen; oc, occipital condyle; ppt, postpalatine torus; pr, promontorium of petrosal.



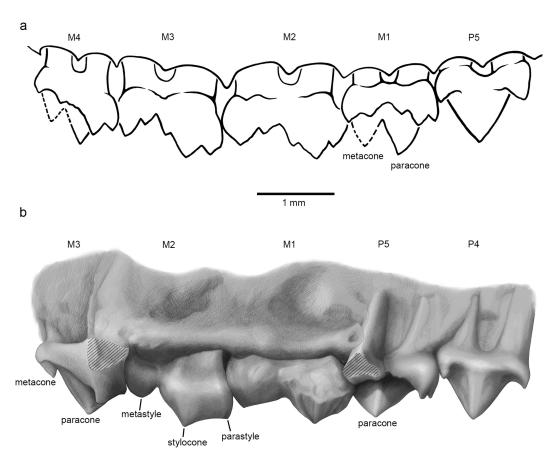
Extended Data Fig. 3 | Left upper and lower teeth of A. zhoui STM33-5 as preserved on the main slab (A) of the specimen from 3D rendering (Mimics) of computed tomography scans. a, Buccal view.

 ${\bf b},$ Lingual view. ${\bf c},$ Occlusal view. The lingual face of $M_1-M_3,$ including the entoconid and hypoconulid, has been sheared off.



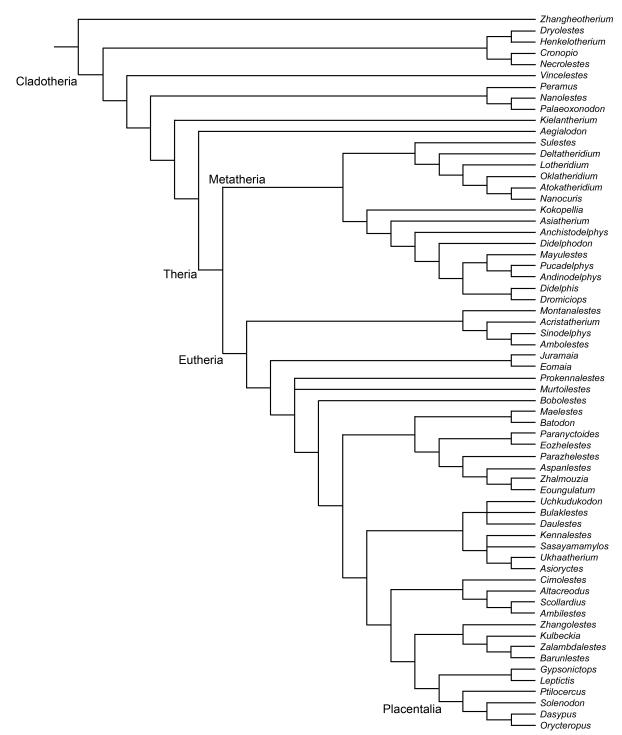
Extended Data Fig. 4 | Comparison of left upper and lower jaws in lateral view of *Ambolestes*, *Sinodelphys*, *Juramaia* and *Eomaia*. Sinodelphys was redrawn from a previous study and reversed from the

original; Juramaia was redrawn from a previous study⁵; and Eomaia was redrawn from a previous study² and reversed from the original. Scale bars, 2 mm.

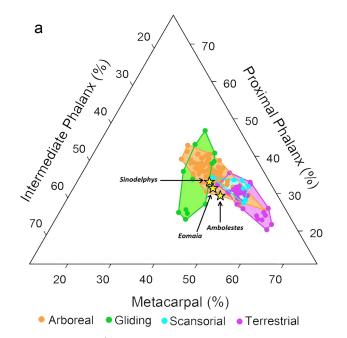


Extended Data Fig. 5 | Left upper dentition of *Sinodelphys szalayi*. a, Dental formula redrawn from a previous study⁷. b, Dental formula proposed in this work on drawing of CM 79002 (a cast of the holotype). Note that the tooth identified as M^1 in the previous study⁷ is not

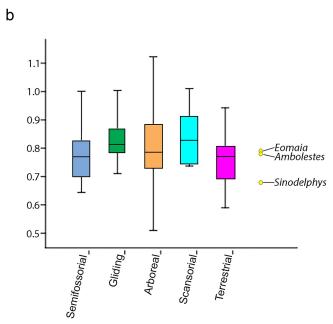
molariform, but is instead built on the same pattern as the tall, trenchant premolariform tooth that is mesial to it. On the M^1 and M^2 (as interpreted here), the paracone and metacone are hidden by the stylar shelf.



Extended Data Fig. 6 | The strict consensus tree of four equally most-parsimonious trees. The consensus tree length = 1,779, consistency index = 0.318 and retention index = 0.597. A simplified version of this consensus tree is presented in Fig. 4.



Extended Data Fig. 7 | Analysis of limb elements of *A. zhoui* STM33-5 for locomotor behaviour. a, Ternary plot showing intrinsic manual ray III proportions. b, Box plots of the intermembral index. The line that divides



the box into two parts represents the median, the box shows the upper and lower quartiles, and the whiskers show extreme values for each group.



Loss of coral reef growth capacity to track future increases in sea level

Chris T. Perry¹*, Lorenzo Alvarez-Filip², Nicholas A. J. Graham³, Peter J. Mumby⁴, Shaun K. Wilson⁵,⁶, Paul S. Kench², Derek P. Manzello⁶, Kyle M. Morgan⁶, Aimee B. A. Slangen¹⁰, Damian P. Thomson¹¹, Fraser Januchowski-Hartley¹², Scott G. Smithers¹³, Robert S. Steneck¹⁴, Renee Carlton¹⁵, Evan N. Edinger¹⁶,¹७, Ian C. Enochs⁶,¹⁰, Nuria Estrada-Saldívar², Michael D. E. Haywood¹⁶, Graham Kolodziej⁶,¹³, Gary N. Murphy¹, Esmeralda Pérez-Cervantes², Adam Suchley², Lauren Valentino⁶,¹³, Robert Boenish²⁰, Margaret Wilson²¹ & Chancey Macdonald²²,²³

Sea-level rise (SLR) is predicted to elevate water depths above coral reefs and to increase coastal wave exposure as ecological degradation limits vertical reef growth, but projections lack data on interactions between local rates of reef growth and sea level rise. Here we calculate the vertical growth potential of more than 200 tropical western Atlantic and Indian Ocean reefs, and compare these against recent and projected rates of SLR under different Representative Concentration Pathway (RCP) scenarios. Although many reefs retain accretion rates close to recent SLR trends, few will have the capacity to track SLR projections under RCP4.5 scenarios without sustained ecological recovery, and under RCP8.5 scenarios most reefs are predicted to experience mean water depth increases of more than 0.5 m by 2100. Coral cover strongly predicts reef capacity to track SLR, but threshold cover levels that will be necessary to prevent submergence are well above those observed on most reefs. Urgent action is thus needed to mitigate climate, sea-level and future ecological changes in order to limit the magnitude of future reef submergence.

SLR will directly impact coastal communities through shoreline inundation and erosion^{1,2}. Along coral-reef-fronted coastlines, the maintenance of reef surface elevation relative to sea level will critically influence magnitudes of future shoreline change and flooding risk^{3,4}. This is because reef structure and water depth modulate across-reef and near-shore wave energy regimes⁵⁻⁷. Mean water depth increases will occur in areas where vertical growth rates lag behind actual or relative (for example, from glacial isostatic adjustment or land subsidence) increases in sea level^{4,8}. This is a widely discussed scenario as the abundance of reef-building species declines globally, limiting reef growth potential $^{9-14}$, while at the same time significant sea-level increases are projected (global mean 0.44 m under RCP2.6 by 2100, 0.74 m under RCP8.5^{15,16}). Even modest depth increases of approximately 0.5 m above reefs are projected to increase coastal flooding risk, and change near-shore sediment dynamics^{3,5,17,18}. However, datasets to support predictions of magnitudes of above-reef submergence and how these may vary geographically under different RCP scenarios are sparse¹⁹. This is a major knowledge gap with important socio-economic and policy implications for urbanized tropical coastlines and reef islands given projected costs of adaptation and mitigation planning⁴.

To estimate reef growth capacity under future SLR, we calculated mean increases in water depth above reefs using a large dataset of reef carbonate budget data collected from more than 200 reefs around two major reef-building regions, the tropical western Atlantic and the Indian Ocean. These data, based on in situ ecological metrics (see Methods), were collected between 2009 and 2017, allowing us to explore intra-regional variations in contemporary carbonate budget states and site-specific temporal dynamics in budget states. Using these data, we derived first-order estimates of maximum vertical reef accretion potential (RAP_{max}, in mm per year (yr⁻¹); see Methods) to explore four key issues. First, we assess inter- and intra-regional variations in site-specific RAP_{max} rates in the context of recent disturbance histories. Second, we use datasets obtained before and after the 2016 bleaching event for impacted Indian Ocean sites to quantify changes in RAP_{max} rates and consider the implications for reef growth given the increasingly important control bleaching has on reef health 14,20,21. Third, we derive best-estimate predictions of reef capacity to track projected rates of SLR, and project total minimum water depth increases at each site by 2100, by comparing site-specific RAP_{max} rates against recent (1993–2010) altimetry-derived regional SLR rates and those projected under RCP4.5 and RCP8.5 scenarios²². Fourth, we quantify the relationship between mean coral cover (as the most widely used reef 'health' metric^{9,10}) and reef submergence under these same SLR scenarios over the next few decades to identify regional coral cover thresholds that are necessary to limit reef submergence.

¹Geography, College of Life and Environmental Sciences, University of Exeter, Exeter, UK. ²Biodiversity and Reef Conservation Laboratory, Unidad Académica de Sistemas Arrecifales, Instituto de Ciencias del Mar y Limnología, Universidad Nacional Autónoma de México, Puerto Morelos, Mexico. ³Lancaster Environment Centre, Lancaster University, Lancaster, UK. ⁴Marine Spatial Ecology Lab, School of Biological Sciences and ARC Centre of Excellence in Coral Reef Science, University of Queensland, Brisbane, Queensland, Australia. ⁵Department of Biodiversity, Conservation and Attractions, Kensington, Perth, Western Australia, Australia. ⁶Oceans Institute, University of Western Australia, Crawley, Western Australia, Australia. ⁷School of Environment, The University of Auckland, New Zealand. ⁸Atlantic Oceanographic and Meteorological Laboratory, NOAA, Miami, FL, USA. ⁹Asian School of the Environment, Nanyang Technological University, Singapore, ¹⁰NIOZ Royal Netherlands Institute for Sea Research, Department of Estuarine and Delta Systems, Utrecht University, Yerseke, The Netherlands. ¹¹CSIRO, Indian Ocean Marine Research Centre, University of Western Australia, Crawley, Western Australia, Australia. ¹²2UMR 248 MARBEC/UMR250 ENTROPIE, UM2-CNRS-IRD-IFREMER-UM1, University of Montpellier 2, Montpellier 2, Montpellier 2, Montpellier 3, Montpellier 3, Montpellier 4, Montpellier 4, Montpellier 4, Montpellier 5, Montpellier 6, Montpellier 7, Montpellier 7, Montpellier 7, Montpellier 8, Montpellier 8, Montpellier 9, Montp

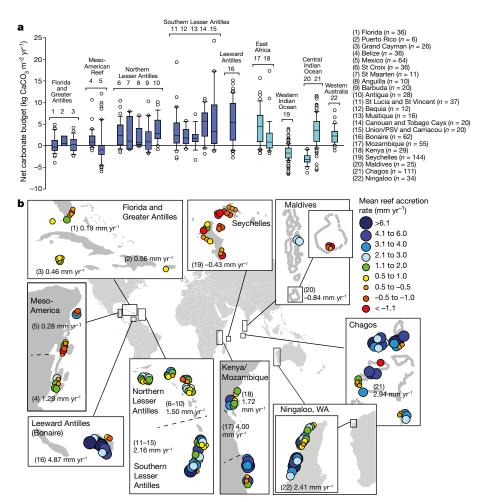


Fig. 1 | Reef carbonate budgets and accretion rates for the tropical western Atlantic and Indian Ocean. a, Plots showing site-level carbonate budget data (kg CaCO₃ m $^{-2}$ yr $^{-1}$) grouped by country or territory within ecoregions. Box plots depict the median (horizontal line), box height depicts first and third quartiles, whiskers represent the 95th percentile, and outliers outside the 95th percentile are shown as circles. Bold numbers

Carbonate budgets and reef accretion potential

Our data show that contemporary carbonate budgets (G, where G = kgCaCO₃ m⁻² yr⁻¹) of most shallow water (less than 10 m depth) reefs across the tropical western Atlantic (2.55 \pm 3.83 G (mean \pm s.d.)) and Indian Ocean $(1.41 \pm 3.02 \,\mathrm{G})$ are currently low (Fig. 1a), and are substantially below the optimal rates (approximately 5–10 G) that have been reported under high coral cover states for both regions²³. Mean carbonate budgets do not differ significantly between the two ocean regions (generalized linear mixed model (GLMM), P = 0.485), but there were significant differences among regions within ocean basins (GLMM, P = 0.046). In the tropical western Atlantic, the highest carbonate budgets were calculated on Leeward Antilles reefs (5.75 \pm 4.87 G; Fig. 1a), a rate that is closer to historical optimal rates²³. The lowest rates were along the Mesoamerican Reef (Mexico, 0.14 ± 3.81 G; Belize, 1.52 ± 2.19 G), in Florida $(0.16 \pm 1.96$ G) and Grand Cayman $(0.28 \pm 1.74 \,\mathrm{G}; \mathrm{Fig.}\ 1a \ \mathrm{and}\ \mathrm{Supplementary}\ \mathrm{Table}\ 1).$ These trends mirror those in coral cover reported in recent basin-wide analyses²⁴, and provide compelling evidence that both coral carbonate production $(4.22\pm4.06\,\mathrm{G})$ and bioerosion rates $(1.74\pm1.46\,\mathrm{G})$ are low across many tropical western Atlantic reefs. As with net G, rates of both coral carbonate production and bioerosion exhibit marked intra-ocean variability (Extended Data Fig. 1) and we note that only sites in the southeast, such as Bonaire (Fig. 1b), are characterized by both high carbonate production and bioerosion rates (8.12 \pm 4.60 G and 2.79 \pm 1.08 G, respectively; see Extended Data Fig. 1) that are close to historically estimated regional rates 19,25. In the Indian Ocean, the highest

above the plots indicate the country or territory. Numbers in italics adjacent to the countries indicate the number of transects per country/territory. \mathbf{b} , Calculated maximum reef accretion potential (RAP_{max}) rates (mm yr⁻¹) for each reef within ecoregions. Numbers in parentheses in each area box denote the country or territory followed by the mean reef accretion rate (mm yr⁻¹).

contemporary budgets were calculated on reefs in Mozambique (4.78 \pm 5.01 G) and Ningaloo, Australia (2.46 \pm 2.01 G). The lowest (and net negative) rates were calculated at the Seychelles (-1.51 ± 1.90 G) and Maldives sites (-2.98 ± 1.30 G; Fig. 1a and Supplementary Table 1).

Low-carbonate budget states are reflected in low calculated RAP_{max} rates at many sites across both oceans. In the tropical western Atlantic, the mean RAP_{max} rate across all sites is $1.87\pm2.16\,\mathrm{mm\,yr^{-1}}$ but there is significant intra-ocean variability (GLMM, $P\!=\!0.032$). The highest RAP_{max} rates were calculated at sites in the southern Lesser Antilles ($2.16\pm1.93\,\mathrm{mm\,yr^{-1}}$) and Leeward Antilles ($4.87\pm2.71\,\mathrm{mm\,yr^{-1}}$; Fig. 1b). Low RAP_{max} rates characterize all reefs examined in Florida and the Greater Antilles (Grand Cayman, $0.46\pm0.66\,\mathrm{mm\,yr^{-1}}$; Florida, $0.19\pm0.93\,\mathrm{mm\,yr^{-1}}$; Fig. 1b) and along the Mesoamerican Reef (Belize, $1.29\pm0.89\,\mathrm{mm\,yr^{-1}}$; Mexico, $0.28\pm1.52\,\mathrm{mm\,yr^{-1}}$; Fig. 1b). These low RAP_{max} rates are likely to result from a prolonged period (at least multi-decadal in duration) of ecological decline driven by various regional-scale factors (fishing pressure, coral disease, bleaching, loss of herbivorous taxa and water quality declines 13,26,) that have substantially changed reef ecology.

In the Indian Ocean, mean calculated regional RAP_{max} rates are only 2.01 ± 2.33 mm yr⁻¹. Sites in East Africa (Mozambique, 4.00 ± 2.78 mm yr⁻¹; Kenya, 1.72 ± 1.32 mm yr⁻¹) and Ningaloo, Australia $(2.41\pm2.01$ mm yr⁻¹) have the highest mean RAP_{max} rates, whereas western and central Indian Ocean sites are on average net negative (Seychelles, -0.43 ± 0.95 mm yr⁻¹; Maldives, -0.84 ± 0.47 mm yr⁻¹; Fig. 1b). This reflects the fact that these areas

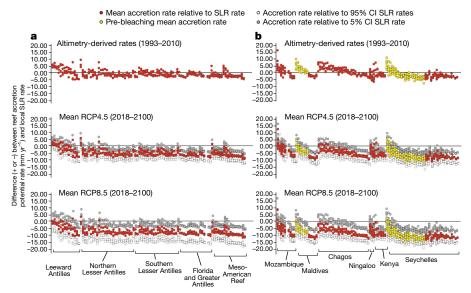


Fig. 2 | Difference between calculated reef accretion potential (mm yr $^{-1}$) relative to recent (1993–2010) and projected rates of SLR. a, b, Plots showing difference between reef accretion rate and SLR for tropical western Atlantic (a; n = 95) and Indian Ocean (b; n = 107) sites. Recent SLR rates are based on altimetry data for the period 1993–2010

(see Methods). Mean RCP4.5 and RCP8.5 SLR rates (and 5% and 95% confidence intervals (CI)) are based on projections for the period 2018–2100³¹ (see Supplementary Table 2). Dots show individual transect data within each site.

were extensively affected by the 2016 bleaching event²⁷ (Extended Data Fig. 2), with widespread coral mortality to depths of at least 6-7 m. Chagos corals also suffered high mortality during 2016²² and although post-event budget assessments have yet to be undertaken it is likely that the relatively high mean RAP_{max} rates that we report $(2.94 \pm 2.06 \,\mathrm{mm\,yr^{-1}}; \,\mathrm{Fig.}\,1)$ for Chagos far exceed contemporary rates. At sites with both pre- and post-2016 data, bleaching significantly reduced both net G (GLMM, P < 0.001) and RAP_{max} (GLMM, P < 0.001). Declines were greatest in the Maldives and on 'recovering reefs'²⁸ in the Seychelles (Extended Data Fig. 2). There were negligible differences on 'regime-shifted' Seychelles reefs as coral cover, net G and accretion were already low. The major consequence of the 2016 event is that most reefs in the impacted areas are presently in net erosional or non-net accretionary states. Furthermore, given: (1) that not all Seychelles reefs recovered successfully from past (1998) bleaching²⁸; and (2) that models predict the rapid onset of annual bleaching for the central Indian Ocean, under both RCP4.5 and RCP8.5 scenarios²¹ (that is, well inside the timescales necessary for reef recovery^{29,30}) the capacity for Indian Ocean reefs to regain high accretion states is increasingly questionable.

Reef accretion and projected SLR

To assess reef capacity to track local SLR, we compared our calculated RAP_{max} rates against recent altimetry-measured SLR rates for the period 1993-2010 (see Methods) and rates projected under RCP4.5 and RCP8.5²² (see Methods and Supplementary Table 2). In both regions only around 45% of reefs have calculated mean RAP_{max} rates close to (within $\pm 1 \,\mathrm{mm}\,\mathrm{yr}^{-1}$) or above local recent (altimetry-derived) SLR rates. Therefore, for many reefs there is already a divergence between reef growth potential and the local recent rate of SLR (Fig. 2). However, these values fall to only 6.2% and 3.1%, respectively, in the tropical western Atlantic, when we compare calculated RAP_{max} rates for each site to projected mean local RCP4.5 and RCP8.5 rates for the twentyfirst century³¹. In the Indian Ocean, only 2.7% of reefs have mean RAP_{max} rates close to (within ± 1 mm yr⁻¹) RCP4.5 projections and 1.3% close to mean RCP8.5 projections (Fig. 2). Although a more positive prognosis would be implied in the Indian Ocean on the basis of pre-bleaching states (59% of the reefs had RAP_{max} rates close to (within ± 1 mm yr⁻¹) recent measured SLR rates; Fig. 2), our data suggest that few reefs in either region will be able to match average twenty-first century projected SLR rates (see Supplementary Table 2) if current ecological conditions persist.

Projections of reef submergence

To assess magnitudes of future reef submergence, we used our calculated RAP_{max} rates to predict total minimum water depth increases above each reef by the end of this century (Fig. 3), and in the Indian Ocean for selected sites based on pre- and post-2016 bleaching data. However, these predictions are probably at the more optimistic end of the spectrum in terms of reef keep-up capacity, both for methodological reasons (see Methods) and because of the lag time between climate warming and SLR. Thus, calculated magnitudes of water depth increase should be considered as best-case scenarios and the minimums for which regions should prepare. Allowing for these caveats, our current projections are that if strong climate mitigation actions can be rapidly implemented (for example, an RCP2.6-type scenario) that restrict SLR rates to close to those measured across our study areas over the last few decades (that is, <3 mm yr⁻¹; see Supplementary Table 2), then the difference between reef accretion and SLR rate will on average be low in both regions, assuming that ecological conditions do not deteriorate further (mean <10 cm increases by 2100; see Supplementary Table 3).

By contrast, significant water depth increases are projected above these reefs by 2100 under both RCP4.5 and RCP8.5 scenarios. Under RCP4.5 projections water depths on the tropical western Atlantic reefs are predicted to increase by 14-66 cm (5-95% confidence interval range) (mean, \sim 40 cm or 4.8 mm yr⁻¹), and between 16 and 104 cm (mean, \sim 60 cm or 7.2 mm yr⁻¹) under RCP8.5 (Fig. 3). In the Indian Ocean mean water depth is estimated to increase by 14-72 cm (mean, $47 \,\mathrm{cm}$ or $5.6 \,\mathrm{mm}$ yr⁻¹) under RCP4.5 and between 22 and 112 cm (mean, 71 cm or 8.5 mm yr⁻¹) under RCP8.5 (Fig. 3 and Supplementary Table 3). Larger average increases of around 63 cm under RCP4.5 $(34\text{--}92\,\text{cm}\,(5\text{--}95\%\,\text{confidence interval range}))$ and $87\,\text{cm}\,(41\text{--}132\,\text{cm})$ under RCP8.5 (Fig. 3 and Supplementary Table 3) are predicted for bleaching-affected central Indian Ocean reefs in the absence of sustained ecological recovery. The major implications are that while 32% of tropical western Atlantic and 45% of Indian Ocean reefs are predicted to experience increases of over 0.5 m by 2100 under mean local RCP4.5 scenarios, under RCP8.5 projections, 80% of our tropical western Atlantic and 78% of Indian Ocean reefs are predicted to experience minimum mean water depth increases above this level. This is an

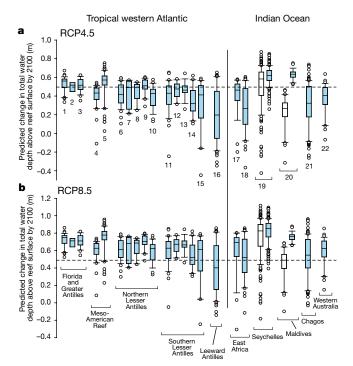


Fig. 3 | Total predicted increases in water depth above reefs by 2100. a, b, Plots for site-level data showing predicted water depth increases against mean RCP4.5 (a) and RCP8.5 (b) SLR projections for the period 2018–2100. Box plots depict median (horizontal line), box height depicts first and third quartiles, whiskers represent the 95th percentile, and outliers outside the 95th percentile are shown as circles. White bars denote pre-bleaching data. The dashed line shows the 0.5-m threshold above which significantly increased wave energy regimes are predicted. Site numbers as in Fig. 1.

important depth threshold as recent models³² suggest that, on average, wave energy regimes will increase especially rapidly once water depth increases exceed 0.5 m. Of major future concern is that, because of the delayed response of processes contributing to SLR (deep ocean warming, and ice sheet and glacier mass loss), these submergence trends are projected to increase towards the end of the century^{16,31,33}. Therefore, the higher end projections of water depth increases for each scenario may be more realistic (Supplementary Table 3), rapidly exacerbating the threat to coastal communities and to small island developing states^{1,4}.

Reef state and submergence trajectories

An especially pressing issue for reef and coastal managers is the question of which reefs are most likely to experience submergence over the coming decades, and how this relates to reef state. The percentage of live coral cover is the most widely reported metric of reef state and we thus used our data to examine whether a metric as simple as coral cover had predictive capacity for projecting changes in sea level above reefs. Although our datasets span two biogeographical provinces, a range of depths (2-13 m) and a diversity of community structures, coral cover explained up to 62% of the projected increase in net water depth by the year 2050 (Fig. 4 and Extended Data Table 1). Simulations uncover that high coral cover states would experience little water depth increase with some even extending closer to the surface. However, statistical fits to our data suggest that coral cover levels of around 40% in the tropical western Atlantic, and approximately 50% in the Indian Ocean, are needed to avoid the prospect of net reef submergence in the next few decades (by 2050) under mean RCP4.5 SLR projections. However, this threshold increases to nearly 60% in the tropical western Atlantic and nearly 70% in the Indian Ocean under the current emissions trajectory of RCP8.5. Given that coral cover levels across the sites in our dataset average only 20.6 \pm 13.9% in the tropical western Atlantic, and $17.8\pm12.6\%$ in the Indian Ocean region (Supplementary Table 1),

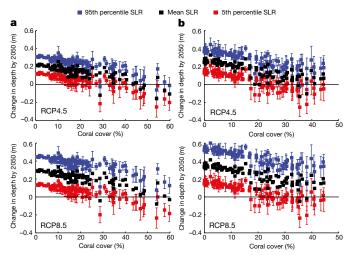


Fig. 4 | Relationships between mean coral cover (%) and changes in water depth (m) above reefs by 2050. a, b, Model simulations (100 per site and SLR scenario) showing predicted changes (y axis) in mean water depth (m) above reefs as a function of coral cover (x axis). a, Tropical Western Atlantic sites (n = 95 reefs). b, Indian Ocean sites (n = 104 reefs). Mean change in depth is shown as the centre point. Error bars are s.d. Simulations show trends under lower (5th percentile), mean and upper (95th percentile) projections of SLR under RCP4.5 and RCP8.5 SLR scenarios.

there is therefore a high probability that mean water depths above reefs will increase by at least a few tens of centimetres in the coming decades.

Summary

The potential for a high proportion of reefs (over 75% across our sites under RCP8.5) to experience water depth increases greater than 0.5 m by 2100 is of concern, because modelling studies suggest this will be sufficient to open higher wave-energy windows that will increase sediment mobility, shoreline change and island overtopping^{1-3,17,18}. We also show that major climate-driven perturbations, specifically coral bleaching, can drive major declines in reef accretion potential. The most worrying end-point scenario is that if predictions of increasing bleaching frequency are realized^{21,34} and result in more frequent mortality, reefs may become locked into permanent low accretion rate states, leading to increasing rates of submergence under all SLR scenarios. Ocean acidification and thermal impacts on calcification represent additional threats and may negatively impact reef calcification and increase bioerosion^{35,36}. These collective threats will be exacerbated by the low coral cover states that define many reefs, and which our analysis suggests will be insufficient to prevent reef submergence. Our approach represents a first step in improving our predictive capabilities in these areas, but given the societal relevance and economic costs of SLR along populated tropical coastlines⁴, and that coral reefs have the potential to have a key role in nature-based defence strategies, these issues should have a high priority on the research agenda.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0194-z.

Received: 2 November 2017; Accepted: 9 May 2018; Published online 14 June 2018.

- Storlazzi, C. D., Elias, E. P. L. & Berkowitz, P. Many atolls may be uninhabitable within decades due to climate change. Sci. Rep. 5, 14546 (2015).
- Kench, P. S., Ford, M. R. & Owen, S. D. Patterns of island change and persistence offer alternate adaptation pathways for atoll nations. *Nat. Commun.* 9, 605 (2018).
- Beetham, E., Kench, P. S. & Popinet, S. Future reef growth can mitigate physical impacts of sea-level rise on Atoll Islands. Earths Future 5, 1002–1014 (2017).

RESEARCH ARTICLE

- Ferrario, F. et al. The effectiveness of coral reefs for coastal hazard risk reduction and adaptation. Nat. Commun. 5, 3794 (2014).
- 5. Baldock, T. E., Golshani, A., Callaghan, D. P., Saunders, M. I. & Mumby, P. J. Impact of sea-level rise and coral mortality on the wave dynamics and wave forces on barrier reefs. Mar. Pollut. Bull. 83, 155–164 (2014).
- Baldock, T. E. et al. Impact of sea-level rise on cross-shore sediment transport on fetch-limited barrier reef island beaches under modal and cyclonic conditions. Mar. Pollut. Bull. 97, 188-198 (2015).
- Quataert, E., Storlazzi, C., van Rooijen, A., Cheriton, O. & van Dongeren, A. The influence of coral reefs and climate change on wave-driven flooding of tropical coastlines. Geophys. Res. Lett. 42, 6407-6415 (2015).
- van Woesik, R., Golbuu, Y. & Roff, G. Keep up or drown: adjustment of western Pacific coral reefs to sea-level rise in the 21st century. R. Soc. Open Sci. 2,
- Bruno, J. F. & Selig, E. R. Regional decline of coral cover in the Indo-Pacific: timing, extent, and subregional comparisons. PLoS ONE 2, e711 (2007).
- 10. Gardner, T. A., Côté, I. M., Gill, J. A., Grant, A. & Watkinson, A. R. Long-term region-wide declines in Caribbean corals. Science **301**, 958–960 (2003).
- 11. Perry, C. T. et al. Caribbean-wide decline in carbonate production threatens coral reef growth. Nat. Commun. 4, 1402 (2013).
- 12. Perry, C. T. et al. Remote coral reefs can sustain high growth potential and may match future sea-level trends. Sci. Rep. 5, 18289 (2015).
- Kennedy, E. V. et al. Avoiding coral reef functional collapse requires local and global action. *Curr. Biol.* 23, 912–918 (2013).
- Hughes, T. P. et al. Global warming and recurrent mass bleaching of corals.
- Nature **543**, 373–377 (2017).

 15. Moss, R. H. et al. The next generation of scenarios for climate change research and assessment. *Nature* **463**, 747–756 (2010).
- Church, J. A. et al. in Climate Change 2013: The Physical Science Basis (ed.
- Stocker, T. F. et al.) Ch. 13 (Cambridge Univ. Press, 2013). 17. Storlazzi, C. D., Elias, E., Field, M. E. & Presto, M. K. Numerical modelling of the impact of sea-level rise on fringing coral reef hydrodynamics and sediment transport. Coral Reefs 30, 83-96 (2011).
- 18. Beetham, E., Kench, P., O'Callaghan, J. & Popinet, S. Wave transformation and shoreline water level on Funafuti Atoll, Tuvalu. J. Geophys. Res. Oceans 121, 311-326 (2016).
- Perry, C. T. et al. Regional-scale dominance of non-framework building corals on Caribbean reefs affects carbonate production and future reef growth. Glob. Change Biol. **21**, 1153–1164 (2015).
- 20. Hoegh-Guldberg, O. Climate change, coral bleaching and the future of the world's coral reefs. Mar. Freshw. Res. 50, 839-866 (1999).
- 21. van Hooidonk, R. et al. Local-scale projections of coral reef futures and implications of the Paris Agreement. Sci. Rep. 6, 39666 (2016).
- 22. Sheppard, C. et al. Bleaching and mortality in the Chagos Archipelago. Atoll Res. Bull. 613, 1-26 (2017).
- 23. Vecsei, A. A new estimate of global reefal carbonate production including the fore-reefs. Glob. Planet. Change 43, 1–18 (2004). 24. Jackson, J. B. C., Donovan, M. K., Cramer, K. L. & Lam, V. V. (eds) Status and
- Trends of Caribbean Coral Reefs: 1970–2012 (Global Coral Reef Monitoring Network, IUCN, Gland, 2014).
- 25. Perry, C. T. et al. Changing dynamics of Caribbean reef carbonate budgets: emergence of reef bioeroders as critical controls on present and future reef growth potential. *Proc. R. Soc. B* **281**, 20142018 (2014).
- 26. Mumby, P. J. & Steneck, R. S. Coral reef management and conservation in light of rapidly evolving ecological paradigms. *Trends Ecol. Evol.* 23, 555–563 (2008)
- 27. Perry, C. T. & Morgan, K. M. Post-bleaching coral community change on southern Maldivian reefs: is there potential for rapid recovery? Coral Reefs 36, 1189-1194 (2017)
- Graham, N. A., Jennings, S., MacNeil, M. A., Mouillot, D. & Wilson, S. K. Predicting climate-driven regime shifts versus rebound potential in coral reefs. Nature **518**, 94-97 (2015).

- 29. Sheppard, C. R. C. et al. Reefs and islands of the Chagos Archipelago, Indian Ocean: why it is the world's largest no-take marine protected area. Aquat. Conserv. 22, 232-261 (2012).
- Pisapia, C. et al. Coral recovery in the central Maldives archipelago since the last major mass-bleaching, in 1998. Sci. Rep. **6**, 34720 (2016). Slangen, A. B. A. et al. Projecting twenty-first century regional sea-level changes.
- Clim. Change 124, 317–332 (2014).
- Siegle, E. & Costa, M. B. Nearshore wave power increase on reef-shaped coasts due to sea-level rise. Earths Future 5, 1054-1065 (2017).
- Carson, M. et al. Coastal sea level changes, observed and projected during the 20th and 21st century. Clim. Change 134, 269–281 (2016).
- Wolff, N. H. et al. Global inequities between polluters and the polluted: climate change impacts on coral reefs. Glob. Change Biol. 21, 3982-3994 (2015).
- 35. Enochs, I. C. et al. Enhanced macroboring and depressed calcification drive net dissolution at high-CO₂ coral reefs. Proc. R. Soc. B 283, 20161742 (2016)
- Schönberg, C. H. L., Fang, J. K. H., Carreiro-Silva, M., Tribollet, A. & Wisshak, M. Bioerosion: the other ocean acidification problem. ICES J. Mar. Sci. 74, 895-925 (2017).

Acknowledgements We thank the many local institutions that supported and facilitated field data collection. Data collection in the tropical western Atlantic was supported through a Leverhulme Trust International Research Network grant (F/00426/G) to C.T.P. and data collection carried out specifically in Mexico was supported through a Royal Society - Newton Advanced Research Fellowship (NA-150360) to L.A.-F. and C.T.P., in Florida and Puerto Rico as part of the National Coral Reef Monitoring Program through NOAA's Coral Reef Conservation Program and Ocean Acidification Program to D.P.M. and in the eastern Caribbean through a National Geographic Research Grant to R.S.S. Data collection in the Indian Ocean was supported in Kenya and Mozambique through a NERC-ESPA-DFID: Ecosystem Services for Poverty Alleviation Programme Grant (NE/K01045X/1) to C.T.P., in the Maldives through a NERC Grant (NE/K003143/1) and a Leverhulme Trust Research Fellowship (RF-2015-152) to C.T.P., in the Chagos Archipelago through a DEFRA Darwin Initiative grant (19-027), in the Seychelles through an Australian Research Council grant (DE130101705) and Royal Society grant (RS-UF140691) to N.A.J.G. and in Ningaloo through the BHP-CSIRO Ningaloo Outlook Marine Research Partnership. P.J.M. acknowledges the Australian Research Council and World Bank/GEF CCRES project for funding. Rebecca Fisher (Australian Institute of Marine Science, Western Australia) provided statistical advice.

Reviewer information Nature thanks I. D. Haigh and I. Kuffner for their contribution to the peer review of this work.

Author contributions C.T.P. conceived the study with support from L.A.-F., N.A.J.G., P.S.K. and K.M.M. C.T.P., N.A.J.G., P.S.K., K.M.M., P.J.M., A.B.A.S. and S.K.W. developed and implemented the analyses. C.T.P. led the manuscript and all other authors contributed data and made substantive contributions to

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0194-7

Supplementary information is available for this paper at https://doi. org/10.1038/s41586-018-0194-z.

Reprints and permissions information is available at http://www.nature.com/

Correspondence and requests for materials should be addressed to C.T.P. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Field data to calculate biological carbonate production and erosion rates, from which net reef carbonate budgets (G, where $G = \text{kg CaCO}_3 \text{ m}^{-2} \text{ yr}^{-1}$) could be calculated, were collected from reef sites spanning both the tropical western Atlantic (TWA) and Indian Ocean regions. All data were collected between 2009 and 2017 (see Supplementary Table 1). At most TWA sites, these data were collected using the ReefBudget methodology³⁷, and for Indian Ocean sites, using a previously reported adapted version of this methodology^{12,38} that factors for regional differences in coral assemblages and bioeroding communities. Data were collected through a number of discrete projects, however, in all cases, the aim was to capture data from the main shallow-water reef-building zones within a range of sites within each country. Survey depths and habitat types thus reflected this variability, although were kept as consistent as possible within countries, and replicate transects within sites were always depth-consistent. In the TWA, data were mostly collected within the 8-10-m depth fore-reef zone. However, where field/logistical conditions allowed, data were collected also at shallower (around 5-m depth) sites, although the number of locations for which data from both depths could be collected was limited. In the Indian Ocean region, survey depths and habitat zones are more variable reflecting the more diverse range of reef types and geomorphologies associated with the countries in our dataset. Our data thus provide an overview of the range in budgetary states, and the resultant accretion potential, of reefs within a country, accepting that not every reef or setting can be realistically assessed. No budget data were collected from high-energy reef-crest settings (<2 m depth) due to physical working constraints, but we note that reported long-term accretion rates for such settings (where these systems are usually dominated by coralline algae) are generally less than 1-2 mm per year^{39,40}, rates that are not dissimilar to those calculated at many sites in this study. The number of replicate transects (see Supplementary Table 1) varied between sites (ranging from 3 to 8) depending on field logistics and weather constraints.

Following the ReefBudget methodology, benthic data were collected using a 10-m transect as a guide line below which a separate 1-m flexible tape was used to measure the distance within each linear 1 m covered by each category of benthic cover. All overhangs, vertical surfaces and horizontal surfaces below the line were thus surveyed. Scleractinian corals were recorded to species level in the TWA, and to genera and morphological level (for example, Acropora branching, Porites massive and so on) in the Indian Ocean. Substrate rugosity was calculated as total reef surface divided by linear distance (a completely flat surface would therefore have a rugosity of 1). To calculate rates of coral carbonate production, we integrated the mean percentage of cover of each coral species with species-specific (or nearest equivalent species) measures of skeletal density (g ${\rm cm}^{-3}$) and linear growth rate (cm yr⁻¹), as derived from published sources (http://geography.exeter.ac.uk/reefbudget/). These data were then combined with rugosity measures to yield a value for coral carbonate production (*G*) relative to actual transect surface area. For several sites in both regions carbonate production rates were calculated slightly differently, because community composition data were based on standard linear intersect methodologies. These were TWA sites in the Windward and Leeward Antilles and, in the Indian Ocean, at Ningaloo and Seychelles. In these cases, individual coral colony cover data were scaled up to derive a three-dimensional measure of cover by using genera or growth form-specific rugosity metrics. For several Indian Ocean sites (Maldives and Seychelles) that were known to be severely affected by the 2016 bleaching event, we also report post-bleaching changes in carbonate production rates, with census data collected using the same methodology as that used pre-bleaching.

To calculate rates of bioerosion, we also undertook census studies to determine abundance and size of parrotfish and bioeroding urchins (both to species level) per unit area of reef following the methods previously reported for TWA and Indian Ocean sites 12,37. All parrotfish abundance data were collected along replicate 30×4 -m² belt transects, except in Chagos (50×5 -m belts), Seychelles (7.5 m radial surveys) and Ningaloo (100×10 -m belts). To calculate bioerosion rates by each individual fish, we used models based on total length and life phase to predict the bite rates (bites per hour) for each species, as reported in Perry et al. 12,37. To calculate bioerosion rates by urchins, we undertook additional surveys at each site, using either 10 \times 2-m or 10 \times 1-m belt transects to determine the species and test sizes of urchins per unit area of reef. Census data were then combined with published species/test class size erosion rate data 12,37 to yield a measure of erosion rate. Rates of endolithic bioerosion were estimated for most TWA sites based on a census of endolithic sponge tissue cover per unit area of reef substrate^{37,41}. Exceptions were sites in Bonaire and the Windward Antilles, where surveys were not conducted and literature-derived rates from the TWA were applied. Endolithic bioerosion rates were estimated at all Indian Ocean sites by applying rates from the literature to available benthic substrate 12 .

To calculate maximum reef accretion potential (RAP_{max}) rates $(mm \ yr^{-1})$ at each site, we followed a previously used method^{11,12} based on the conversion of measured site-specific net carbonate production rates (G) as proposed previously⁴².

In this conversion net carbonate production is taken as the sum of calculated gross carbonate production by corals and coralline algae minus erosion rate. We then also factored for variations in accumulating reef framework porosity as a function of coral community type and for sediment reincorporation⁴². Stacking porosity values ranging from around 80% void space for branching coral assemblages to about 20% for head coral-dominated assemblages, with rates of approximately 50% for mixed assemblages, were proposed previously 43 . However, since coral communities are rarely entirely monospecific, we used the following assumptions in our calculations: that void space estimates of 30% were appropriate for head and massive coral-dominated assemblages, 70% for branched and tabular coraldominated assemblages and 50% for mixed coral assemblages as determined for each site from benthic coral community data. Sediment reincorporation was factored for by allowing for a proportion of the bioeroded framework (that is converted to sediment) to be reincorporated back into the accumulating reef structure. This proportion was calculated as the sum of 50% of the parrotfish-derived sediment (as a highly mobile bioeroder that defecates randomly over the reef), as well as all sediment produced by urchins and by macrobioerosion. To keep our estimates conservative, we worked on the assumption that only around 50% of this bioerosional sediment yield is actually incorporated back into the reef (see also Hubbard et al.⁴⁴), and excluded any sediment generation by other benthic sediment producers (for example, Halimeda).

Owing to the absence of empirical data on rates of physical reef framework removal per unit area of reef surface over time, we did not factor for physical loss rates. For the same reason, we also did not factor for chemical dissolution of the substrate. The accretion rates that we report, which we consider as current best-estimates of accretion potential across the entire upper portion of a reef profile (on the basis that accretion can result from both in situ coral accumulation and the supply of physically derived rubble from shallow fore-reef areas to the crest/flat 45 , are thus defined as a rate of maximum reef accretion potential, or RAP_{max}). We therefore consider these rates to represent the upper limits of how fast reefs may be accreting at present, and acknowledge that if physical framework loss and chemical dissolution rates 46 could be appropriately factored for at the site level our projected rates would probably be lower. How much lower will depend on spatial variations in physical disturbance regimes and the susceptibility of coral taxa to physical disturbances, and both are likely to vary markedly at intra-regional scales. Testing the validity of our high end (RAP_{max}) rates is thus not simple.

Evidence from Holocene core records of reef growth, when ecological conditions (in terms of the abundance of high-rate carbonate-producing taxa, for example, *Acropora* spp.) are considered to have been more optimal, suggest that many reefs exhibited an impressive capacity to either 'keep-up' or to 'catch-up' during periods of past rapid SLR. Indeed, calculated vertical accretion rates from the early Holocene, when sea levels were rising rapidly, were as high as 12–15 mm yr⁻¹ in both the TWA and Indian Ocean regions⁴⁷. Although longer term average accretion rates were lower (for example, approximately 3–4 mm yr⁻¹ in the TWA ⁴⁸; and a little below this in the Indian Ocean region⁴⁷), these still exceed those estimated for many modern reefs in our dataset, and fall well below even mean RCP4.5 SLR scenarios (see Supplementary Table 2). Furthermore, reef core studies that might allow some assessment of very recent accretion histories on a site-by-site basis, that is with a focus on the last couple of hundred years of reef growth, are sparse/ absent and would make for inherently problematic comparisons because of the magnitudes of coral community change that have occurred at most sites over the last few decades.

However, one useful (albeit subarea-specific) comparator is the recent work of Yates et al. 49 , which used historical bathymetric data from the 1930s to 1980s and Lidar-derived digital elevation models from the late 1990s to 2000s in Florida to calculate net changes in seafloor elevation. This data integrates for the effects of any physical and chemical losses and suggests net negative accretion rates in the upper Florida Keys of around $-1.5\,\mathrm{mm}\,\mathrm{yr}^{-1}$ (over the past 68 years), of $-4.5\,\mathrm{mm}\,\mathrm{yr}^{-1}$ in the lower Florida Keys (over the past 66 years) and of $-2.7\,\mathrm{mm}\,\mathrm{yr}^{-1}$ in the US Virgin Islands (over the past 33 years). Our data from different sites in this region (southeast Florida, the upper Florida Keys and the Dry Tortugas) and which do not include data from lagoon sands and seagrass beds that were integrated within the previous study 49 , have average contemporary accretion rates of -0.4, 1.7 and 0.8 mm yr $^{-1}$, respectively. Our rates are thus, as expected for the various reasons outlined above, a little but not markedly higher, suggesting they provide a reasonable estimate of high end reef accretion potential.

To test for differences in net G and calculated accretion rates between sites and countries across our dataset, we fitted GLMMs to assess whether rates showed statistically significant differences between oceans and regions (n = 885 transects), as well as for the effects of bleaching and the interaction with location (Maldives, Seychelles recovering and regime-shifted) (n = 338 transects), while controlling for site depth and the random effect of site. All GLMMs were fitted using a Gaussian distribution via the lmer function of the package lme4⁵⁰ in R⁵¹, with significance assessed using F-ratio statistics calculated via the ANOVA function in the CAR⁵²

package. Model assumptions of normality and homogeneity of variance were assessed graphically and found to be adequately met. We found a very weak effect of depth on net G (and thus RAP $_{\rm max}$ rates) across our dataset, with net G typically being slightly higher on the deeper reefs ($P\!=\!0.001, r\!=\!0.160$). Although our datasets do not allow a detailed consideration of this issue at the within-region level, the fact that average accretion rates do not noticeably decline with depth across the upper fore-reef depth intervals is consistent with trends inferred from Holocene core records in the TWA region⁴⁸.

To assess the capacity of the reefs in our datasets to match recently observed and future projected changes in sea level, and to estimate magnitudes of water depth increases relative to projected reef accretion by 2100 at each site, we compared our calculated RAP_{max} data against local sea-level change data (Supplementary Table 2). In these comparisons, we assume steady-state ecological conditions persisting. For recent observed rates of change, we compared our RAP_{max} rates against altimetry data for the period 1993-2010 from combined TOPEX/Poseidon, Jason-1, Jason-2/ OSTM and Jason-3 satellite altimetry fields (http://www.cmar.csiro.au/sealevel/ sl_data_cmar.html; downloaded on 22 January 2018). The fields used are monthly averages on a 1° \times 1° grid with the seasonal (annual and semi-annual) signal removed, and include inverse barometer and GIA corrections. The observed rates were computed by fitting a linear trend to the monthly 1993-2010 time series at the nearest available ocean grid point to the reef location. For the period 2018-2100, we used sea-level projections under the RCP4.5 and RCP8.5 scenarios 31,33 . These regional sea-level projections factor for changes in ocean density and dynamics, changes in atmospheric pressure, and glacier and ice sheet surface mass balance contributions based on output from 21 CMIP5 (Climate Model Intercomparison Phase 5⁵³) atmosphere-ocean coupled climate models. In addition, the projections account for model-based contributions from anthropogenic groundwater extraction, for glacial isostatic adjustment and observation-based estimates of ice sheet dynamical processes. The regional sea-level patterns of mass redistribution account for changes in gravitational, deformational and rotational feedbacks. As for the recent observed rates of change, the spatial resolution of the SLR projections is $1^{\circ} \times 1^{\circ}$ and the closest grid point (nearest neighbour) is extracted for comparison to the coral reef data (Supplementary Table 2).

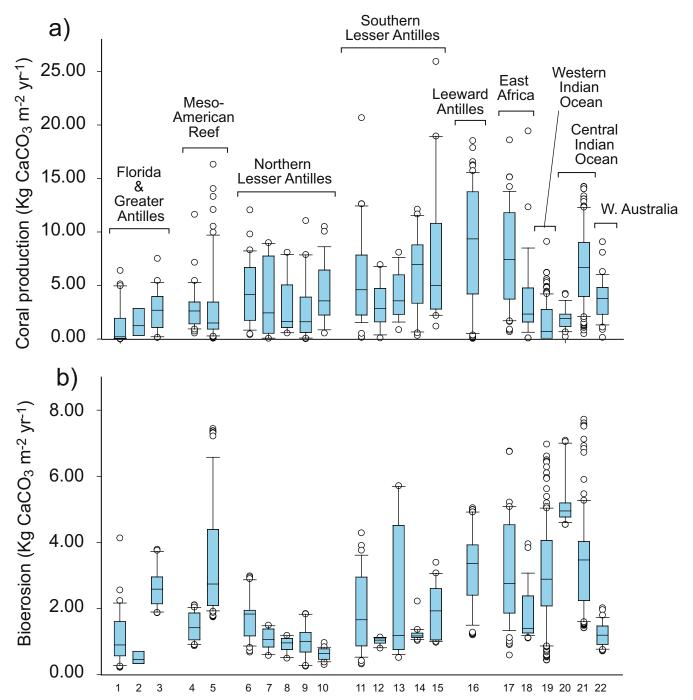
To obtain a greater insight into the importance of coral cover on near-future reef submersion, we undertook Monte Carlo simulations of carbonate budgets, potential accretion rates and projected increases in depth under SLR. One hundred simulations were carried out per site during which community structure was sampled randomly from the site-level statistical distribution of corals, CCAs and sources of bioerosion (that is, sampling from the observed mean and standard deviation of species-specific *G* or erosion rate at the site). Each simulation was extended to estimate the change in seawater depth at the year 2050 for six reference rates of SLR (as above): the 5th percentile, mean, and 95th percentile of the rate of SLR for each of two greenhouse gas (GHG) emission scenarios, RCP4.5 and RCP8.5. For each site, we obtained the mean and standard deviation for each of the six SLR references. Analyses of differences in accretion rate, rates of SLR, and increases in depth over reefs were carried out using non-parametric mixed effects models based on Euclidean distance⁵⁴. This technique is analogous to parametric linear mixed effects models but makes no assumptions about the statistical distribution of errors. Fixed effects included biogeographical region (TWA versus Indian Ocean), GHG emissions scenario (RCP4.5 versus RCP8.5) and coral cover. Country was added as a random effect nested within biogeographical region. The only exception to this approach was the use of linear mixed effects models in order to estimate threshold levels of coral cover where the net submergence of reefs was zero. Models were fitted using the same structure as in PERMANOVA⁵⁵ but the predict function was used to estimate model fits for y = 0. Analysis showed that a shift towards lower GHG emissions (RCP4.5) reduced the degree of reef submergence (Fig. 4; PERMANOVA, P < 0.001) and emissions scenario gained in importance when switching from lower to mean to upper (95 percentile) bounds of projected SLR, explaining 2%, 44%, and 54% of the variance in reef submergence, respectively (Extended Data Table 1). Under the upper bounds of SLR, biogeographical region also became significant (PERMANOVA, P = 0.005) with submergence being slightly greater in the Indian Ocean (Fig. 4b). Under this pessimistic scenario, threshold levels of coral cover required to avoid net reef submergence

were approximately 13% higher in the Indian Ocean than the TWA (73% versus 60%) even under RCP4.5. This relative vulnerability of reefs in the Indian Ocean was associated with higher rates of SLR (0.94 mm yr $^{-1}$ greater; PERMANOVA, $P\!=\!0.02$, Extended Data Tables 2, 3) rather than any biogeographical difference in accretion potential (PERMANOVA, $P\!=\!0.65$; Extended Data Table 4). Although Indian Ocean reefs are generally more resilient than those of the TWA 56 , current ecological trajectories suggest that few coral reef locations will be likely to maintain sufficiently high coral cover levels to keep pace with future SLR, resulting in greater incident wave energy exposure, and changing spectrum of wave processes, along reef-fronted shorelines $^{3.6}$.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

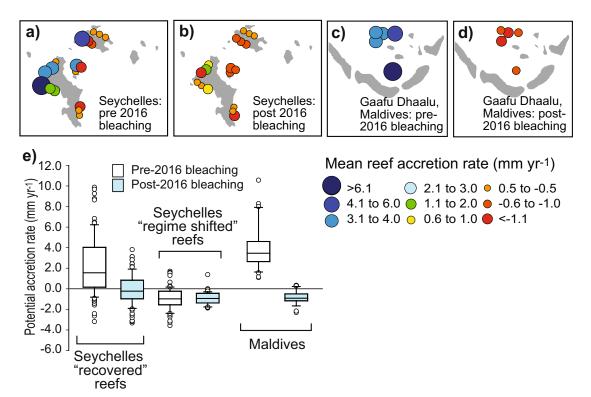
Data availability. Net carbonate budget and reef accretion rate data, and measured and projected sea-level data supporting the findings of this study are available within the paper and its Supplementary Information. Site-level coral cover and carbonate production and bioerosion datasets are available from the corresponding author upon request.

- Perry, C. T. et al. Estimating rates of biologically driven coral reef framework production and erosion: a new census-based carbonate budget methodology and applications to the reefs of Bonaire. *Coral Reefs* 31, 853–868 (2012).
- Januchowski-Hartley, F. A., Graham, N. A. J., Wilson, S. K., Jennings, S. & Perry, C. T. Drivers and predictions of coral reef carbonate budget trajectories. *Proc. R.* Soc. B 284, 20162533 (2017).
- Steneck, R. S., Macintyre, I. G. & Reid, R. P. A unique algal ridge system in Exuma Cays, Bahamas. Coral Reefs 16, 29–37 (1997).
- Gherardi, D. F. M. & Bosence, D. W. J. Late Holocene reef growth and relative sea-level changes in Atol das Rocas, equatorial south Atlantic. *Coral Reefs* 24, 264–272 (2005).
- Murphy, G. N., Perry, C. T., Chin, P. & McCoy, C. New approaches to quantifying bioerosion by endolithic sponge populations: applications to the coral reefs of Grand Cayman. Coral Reefs 35, 1109–1121 (2016).
- Smith, S. V. & Kinsey, D. W. Calcium carbonate production, coral reef growth, and sea level change. Science 194, 937–939 (1976).
- Kinsey, D. W. & Hopley, D. The significance of coral reefs as global carbon sink—response to greenhouse. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 89, 363–377 (1991).
- Hubbard, D. K., Miller, A. I. & Scaturo, D. Production and cycling of calcium carbonate in a shelf-edge reef system (St. Croix, U.S. Virgin Islands): applications to the nature of reef systems in the fossil record. *J. Sedim. Petrol.* 60, 335–360 (1990).
- Blanchon, P. et al. Retrograde accretion of a Caribbean fringing reef controlled by hurricanes and sea-level rise. Front. Earth Sci. 5, 78 (2017).
- Eyre, B. D., Andersson, A. J. & Cryonak, T. Benthic coral reef calcium carbonate dissolution in an acidifying ocean. *Nat. Clim. Change* 4, 969–976 (2014).
- Dullo, W. C. Coral growth and reef growth: a brief review. Facies 51, 33–48 (2005).
- Hubbard, D. K. Depth- and species-related patterns of Holocene reef accretion in the Caribbean and western Atlantic: a critical assessment of existing models. *Int. Assoc. Sedimentol. Spec. Publ.* 41, 1–18 (2009).
- Yates, K. K., Zawada, D. G., Smiley, N. A. & Tiling-Ránge, G. Divergence of seafloor elevation and sea level rise in coral reef ecosystems. *Biogeosciences* 14, 1739–1772 (2017).
- 50. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using Ime4. *J. Stat. Softw.* **67**, 1–48 (2015).
- R Core Team. R: A language and Environment for Statistical Computing https:// www.R-project.org/ (R Foundation for Statistical Computing, Vienna, Austria, 2017).
- Fox, J. & Weisberg, S. An R Companion to Applied Regression 2nd edn (Sage, Thousand Oaks, 2011).
- 53. Taylor, K., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).
- Anderson, M. J. A new method for non-parametric multivariate analysis of variance. Austral Ecol. 26, 32–46 (2001).
- Pinheiro, J. C. & Bates, D. M. Mixed-effects Models in S and S-plus (Springer-Verlag, New York, 2000).
- Roff, G. & Mumby, P. J. Global disparity in the resilience of coral reefs. Trends Ecol. Evol. 27, 404–413 (2012).



Extended Data Fig. 1 | TWA and Indian Ocean coral carbonate production and bioerosion rates. Plots showing mean site level coral carbonate production rate (a) and bioerosion rate (b) data (kg CaCO₃ m⁻² yr⁻¹) grouped by country or territory within ecoregions for TWA and Indian Ocean sites. Box plots depict the median (horizontal line), box height depicts first and third quartiles, whiskers represent the 95th percentile, and outliers outside the 95th percentile are shown as circles. Country/territory codes are as follows: (1) Florida (n = 36); (2) Puerto Rico (n = 6); (3) Grand Cayman (n = 26); (4) Belize (n = 36); (5) Mexico

(n=64); (6) St. Croix (n=36); (7) St. Maarten (n=11); (8) Anguilla (n=10); (9) Barbuda (n=20); (10) Antigua (n=28); (11) St. Lucia and St. Vincent (n=37); (12) Bequia (n=12); (13) Mustique (n=16); (14) Canouan and Tobago Cays (n=20); (15) Union/PSV and Carriacou (n=20); (16) Bonaire (n=62); (17) Mozambique (n=55); (18) Kenya (n=29); (19) Seychelles (n=144); (20) Maldives (n=25); (21) Chagos (n=111); (22) Ningaloo (n=34). n indicates the number of transects per country or territory.



Extended Data Fig. 2 | Reef accretion before and after the central Indian Ocean 2016 bleaching event. a–d, Calculated RAP $_{\rm max}$ rates (mm yr $^{-1}$) before (a, c) and after (b, d) the 2016 bleaching event in the Seychelles and the Maldives. e, Plot shows changes in RAP $_{\rm max}$ rates at 'recovered' (n=96) and 'regime-shifted' reefs 37 (n=72 pre-bleaching, n=48

post-bleaching) in the Seychelles, and Maldives (n=35 pre-bleaching, n=25 post bleaching). Box plots depict the median (horizontal line), box height depicts first and third quartiles, whiskers represent the 95th percentile, and outliers outside the 95th percentile are shown as circles.

Extended Data Table 1 \mid Effects of biogeography, coral cover, GHG emissions scenario and range of SLR projection on the future submergence of coral reefs by 2050

A) Future submergence (depth change) based on lower (5th percentile) SLR projections

						Variance	
Source	df	SS	MS	Pseudo-F	P(perm)	્ર	
coralcover	1	1.9178	1.9178	135.76	0.001	62.1	
Region	1	1.1627E-3	1.1627E-3	0.11133	0.725	2.0	
RCP	1	2.8203E-2	2.8203E-2	15.608	0.001	1.7	
Country (Region)	23	0.31843	1.3845E-2	7.662	0.001	11.3	
Res	363	0.65591	1.8069E-3			22.9	
Total 389	2.921	5					

B) Future submergence (depth change) based on mean SLR projections

						Variance	
Source	df	SS	MS	Pseudo-F	P(perm)	olo	
coralcover	1	1.9866	1.9866	140.62	0.001	44.6	
Region	1	5.4554E-2	5.4554E-2	1.7012	0.197	1.1	
RCP	1	0.65409	0.65409	361.04	0.001	30.5	
Country (Region)	23	0.31843	1.3845E-2	7.6419	0.001	7.9	
Res	363	0.65764	1.8117E-3			15.9	
Total 389	3.671	3					

C) Future submergence (depth change) based on upper (95th percentile) SLR projections

						Variance	
Source	df	SS	MS	Pseudo-F	P(perm)	00	
coralcover	1	2.0557	2.0557	145.51	0.001	26.5	
Region	1	0.24799	0.24799	7.4611	0.005	5.7	
RCP	1	2.1014	2.1014	1151.9	0.001	54.2	
Country(Region)	23	0.31843	1.3845E-2	7.5891	0.001	4.5	
Res	363	0.66222	1.8243E-3			9.1	
Total 389 5.	3857						

Results of PERMANOVA analyses with coral cover, biogeographic region (TWA versus Indian Ocean) and GHG emissions scenario (RCP4.5 versus RCP8.5) as fixed effects and country nested within (biogeographic) region as random effect.



Extended Data Table 2 | Effect of biogeographic region on rates of SLR

						Variance	
Source	df	SS	MS	Pseudo-F	P(perm)	ଚ	
Region	1	4.0348	4.0348	5.5794	0.02	33.9	
Country(Region)	25	20.489	0.81955	835.79	0.001	65.6	
Res	168	0.16474	9.8057E-4			0.5	
Total 194 39	115						

 $PERMANOVA\ analysis\ testing\ the\ effect\ of\ biogeographic\ region\ on\ the\ upper\ 95\%\ of\ predicted\ rates\ of\ SLR.$



Extended Data Table 3 | Differences between SLR rates between biogeographic regions (mm yr⁻¹)

	Diff	Difference in SLR projection (mm yr ⁻¹)					
	(India	(Indian Ocean - Tropical Western Atlantic)					
		Component of SLR Projection					
	Lower bound (5th percentile)	Mean	Upper bound (95 th				
			percentile)				
RCP4.5	0.03	0.33	0.61				
RCP8.5	0.59	0.76	0.94				

The difference in SLR rates between biogeographic regions (mm yr^{-1}) under two GHG emission scenarios and for all three components of SLR projections. Projections are higher in the Indian Ocean except in RCP4.5 lower percentile (0.03), which was not significant.



Extended Data Table 4 \mid Variability in potential accretion rate

						Variance	
Source	df	SS	MS	Pseudo-F	P(perm)	용	
coralcover	1	1756.9	1756.9	135.43	0.001	63.2	
Region	1	2.4775	2.4775	0.15683	0.653	1.8	
Country(Region)	23	292.4	12.713	7.6839	0.001	11.6	
Res	364	602.25	1.6545			23.4	
Total 389 2654	4.1						

Results of PERMANOVA analysis showing local (coral cover) versus regional (TWA versus Indian Ocean) effects on the variability in potential accretion rate.



Structural basis of mitochondrial receptor binding and constriction by DRP1

Raghav Kalia^{1,2,3}, Ray Yu-Ruei Wang^{1,3,4}, Ali Yusuf^{1,3}, Paul V. Thomas^{1,3}, David A. Agard^{1,3,4}, Janet M. Shaw^{2,5} & Adam Frost^{1,2,3,6}*

Mitochondrial inheritance, genome maintenance and metabolic adaptation depend on organelle fission by dynamin-related protein 1 (DRP1) and its mitochondrial receptors. DRP1 receptors include the paralogues mitochondrial dynamics proteins of 49 and 51 kDa (MID49 and MID51) and mitochondrial fission factor (MFF); however, the mechanisms by which these proteins recruit and regulate DRP1 are unknown. Here we present a cryo-electron microscopy structure of full-length human DRP1 co-assembled with MID49 and an analysis of structure- and disease-based mutations. We report that GTP induces a marked elongation and rotation of the GTPase domain, bundle-signalling element and connecting hinge loops of DRP1. In this conformation, a network of multivalent interactions promotes the polymerization of a linear DRP1 filament with MID49 or MID51. After co-assembly, GTP hydrolysis and exchange lead to MID receptor dissociation, filament shortening and curling of DRP1 oligomers into constricted and closed rings. Together, these views of full-length, receptor- and nucleotide-bound conformations reveal how DRP1 performs mechanical work through nucleotide-driven allostery.

Fragmentation of the mitochondrial reticulum disperses units of the organelle during cell division^{1,2}, coordinates morphological adaptation with metabolic demand^{3,4} and quarantines damaged units for turnover^{5,6}. Recent work has also led to the discovery of the role of mitochondrial fission in regulated cell death pathways^{7–9}, brain development and synaptic function^{10,11}, and how certain pathogens disrupt these processes and hijack mitochondrial resources^{12,13}. There is a growing understanding of how inter-organelle contacts between the endoplasmic reticulum and mitochondria initiate mitochondrial fission^{14,15}, and how this process impacts mitochondrial genome duplication and integrity^{16,17}. The master regulator that unites these processes across eukaryotic evolution is the membrane-remodelling GTPase DRP1^{2,18,19}.

DRP1 is necessary, but not sufficient, for mitochondrial fission because receptors must recruit the enzyme to the outer mitochondrial membrane. In mammals, these receptors include MFF and the paralogues MID49 and MID51^{20–24}. After receptor-dependent recruitment, DRP1 assembles into polymers that encircle mitochondria and, by poorly understood mechanisms, channels energy from GTP binding, hydrolysis and nucleotide exchange into a mechanochemical constriction^{8,23,25–28}. In addition to DRP1 and its outer mitochondrial membrane receptors, a recent study revealed that a second member of the dynamin family of GTPases, dynamin-2, may enact the final fission event downstream of DRP1-driven constriction of a mitochondrial tubule²⁹. As such, mitochondrial division is a stepwise reaction regulated by DRP1 receptor binding, oligomerization and guanine nucleotide-dependent conformational dynamics.

We and others have reported that the outer mitochondrial membrane receptors MFF, MID49 and MID51 are independently sufficient to recruit DRP1 to divide mitochondria^{20,21,23,30}. Previous work indicated that the mitochondrial receptor in yeast, Mdv1p, can co-assemble with the DRP1 homologue Dnm1p³¹. We observed that MID49 co-assembled with DRP1 to form a copolymer with altered properties compared with DRP1-only oligomers²³. Although these results suggest that organelle receptors could nucleate and alter the architecture of a

dynamin polymer, the organization and functions of such a co-assembly in organelle constriction remain unclear.

Here we report structural snapshots of DRP1-driven constriction through a MID49- or MID51-dependent reaction. We used cryoelectron microscopy (cryo-EM) to observe how nucleotide binding to the GTPase (G) domain induces conformational changes that allosterically propagate through the bundle-signalling element (BSE) to open and elongate DRP1 and expose multiple surfaces required for receptorbinding and polymerization. 3D reconstruction revealed how the binding of MID49 or MID51 stabilized an arrangement of GTP-bound DRP1 tetramers and nucleated polymerization of a linear cofilament. Next, we reconstituted a path-dependent reaction to observe how GTP hydrolysis and nucleotide exchange lead to conformational constriction by the DRP1 polymer. Specifically, when DRP1 subunits within the MID49 cofilament were allowed to exchange and hydrolyse GTP, they released the receptor and the polymers shortened while curling into closed rings. Finally, we designed phosphomimetic, structure-based and disease-causing mutations to validate our structural models and to examine the allosteric rearrangements that govern curling of linear strings into closed and constricted rings after receptor dissociation.

So far, many structural studies of dynamin-family proteins have relied upon mutated or truncated constructs to facilitate crystallization. We purified wild-type, full-length human DRP1 including the N-terminal G domain, BSE and the four-helix bundle known as the stalk (Fig. 1a). This construct also contained the lipid-binding region of around 100 amino acids that is known as the variable domain. This domain resides between the third and fourth α -helices of the stalk, analogous to the pleckstrin homology domain that is found in endocytic dynamin proteins. A crystal structure of a nucleotide-free and a truncated DRP1 mutant revealed the organization of these domains and an overall similarity to the structure of nucleotide-free endocytic dynamin 25,32,33 . We also purified soluble truncations of MID49 and MID51 that were engineered to lack their N-terminal transmembrane anchors but include the cytoplasmic nucleotidyltransferase-like

¹Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA. ²Department of Biochemistry, University of Utah, Salt Lake City, UT, USA. ³California Institute for Quantitative Biomedical Research, San Francisco, CA, USA. ⁴Howard Hughes Medical Institute, San Francisco, CA, USA. ⁵Howard Hughes Medical Institute, Salt Lake City, UT, USA. ⁶Chan Zuckerberg Biohub, San Francisco, CA, USA. *e-mail: adam.frost@ucsf.edu

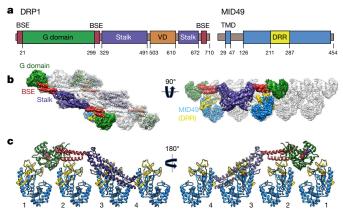


Fig. 1 | Architecture of the DRP1-MID49 linear filament. a, Domain arrangements of DRP1 (left) and MID49 (right). b, Density map and atomic models of DRP1 and MID49(126-454). Green, G domain; Red, BSE; purple, stalk; blue, MID49; yellow, DRR of MID49. c, Each DRP1 chain contacts four different MID49 molecules through receptor interfaces 1-4. as numbered.

domain and the dynamin recruitment region (DRR) required for DRP1 binding $^{20,34-37}$ (Fig. 1a).

Structure of the DRP1-MID49 cofilament

Cofilament assembly resulted upon incubating equimolar ratios of DRP1 with soluble MID49(126-454), MID51(132-463), or both proteins together, in the presence of Mg²⁺, GTP or the GTP analogues GMPPCP or GTP γ S, but not in the presence of other nucleotides (Extended Data Fig. 1). We focused on the filaments formed with MID49(126-454) in the presence of GMPPCP, and determined their structure from cryo-EM images to an average resolution of 4.2 Å (from around 3.5 Å to around 8 Å, Extended Data Figs. 2–6 and Supplementary Table 1). 3D reconstruction revealed a polymer comprising three equivalent faces that meet through defined vertices to form a triangular assembly (Extended Data Figs. 2, 3). A combination of helical reconstruction, segmentation, and single-particle alignment and averaging resolved the elongated DRP1 subunits bound stoichiometrically to MID49(126-454), but no density was assignable for the majority of the variable domain (Fig. 1b, Extended Data Figs. 2–6). Surprisingly, each chain of DRP1 bound MID49 through four different surfaces, and each MID49 in turn bound four DRP1 molecules to yield a vast interaction network (Fig. 1b,c, Extended Data Figs. 2, 3a-c). MID49 binding to four DRP1 molecules stabilized a linear arrangement of inter-DRP1 interfaces, similar to those observed for other dynamin-family proteins^{25,32,33,38–40} (Fig. 1b, Extended Data Figs. 2c, 3d). We refer to the four distinct surfaces of DRP1 that contribute to MID49 and MID51 binding as receptor interfaces 1 to 4 (Fig. 1c).

Structure-based mutants disrupt DRP1-MID49 assembly

The DRR motif of MID49 occupied the space between two neighbouring G domains and contacted both via receptor interfaces 1 and 2 (buried surface areas of around 530 Ų and around 200 Ų, respectively, Figs. 1b,c, 2a). The precise spacing required for this bivalent G-domain interaction explains why previous mutagenesis efforts suggested that the size and topology of the $\beta4-\alpha4$ loop, rather than its exact sequence (which differs between MID49 and MID51), are critical determinants of binding 34,35,37 . Nevertheless, the structure indicated that R235 of MID49 (analogous to R243 in MID51, Extended Data Fig. 7) within the DRR makes key contacts between two neighbouring G domains (Fig. 2a). Accordingly, the MID49(R235E) point mutant could not co-assemble with DRP1 (Fig. 2d–f). In addition, we mutated conserved DRP1 residues involved in receptor interface 1, which is the largest interaction interface. Both the D190A mutation, which should neutralize a salt bridge with the receptor, and the D221A mutation, which

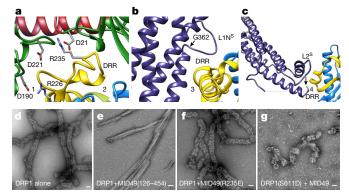


Fig. 2 | Key DRP1-MID49 receptor interfaces and regulatory phosphorylation site. a, Receptor interfaces 1 and 2. Green ribbons on either side of the DRR come from two separate G domains. The residues involved in key interactions of the interfaces are shown as sticks. b, Rotated view of receptor interface 3. Disease-associated DRP1 residue G362 (indicated by the arrow) supports the conformation of the L1N^s loop that is essential for linear copolymerization with MID49. c, Receptor interface 4 with MID49. The arrow points to the DRP1 phosphorylation site S611, at the variable domain-stalk junction. The dashed line indicates the unresolved amino acids of the variable domain. d-g, Negative stain micrographs showing assemblies of DRP1 alone (d), DRP1 with wild-type MID49(126-454) (e), DRP1 with the MID49(126-454) mutant R235E (f; these assemblies resemble those of DRP1 alone), the DRP1 mutant S611D with MID49(126-454) (g; these assemblies also resemble those of DRP1 alone). Scale bars, 50 nm.

should alter the conformation of a key loop within receptor interface 1, prevented co-assembly with MID49 (Fig. 2a). These substitutions also altered the self-assembly properties of DRP1, which suggests pleiotropic effects on nucleotide handling and receptor binding⁴¹ (Extended Data Fig. 8). Specifically, D190 and D221 are also involved in nucleotide-dependent G-domain dimerization (PDB: 3W6O), which leads to the possibility that MID49 and MID51 may modulate GTP-dependent conformational dynamics.

Unexpectedly, MID49 also made contact with the stalk loops of a third and fourth DRP1 molecule through receptor interfaces 3 and 4 (buried surface areas of around 450 Å² and around 230 Å², respectively, Figs. 1c, 2b,c). The DRP1 loops involved in these receptor interfaces, the L1N^S and L2^S loops, are key determinants of assembly for other dynamin-family oligomers^{25,32,33,38,40}. Receptor interface 3 in particular involves the conserved loop L1NS, and is the site of several disease alleles that correlate with elongated mitochondrial morphologies^{42–44}, including G362D and G363D (Fig. 2b, Extended Data Figs. 6b, e, 7a). Previous work has established that this loop comprises part of the binding site for the pleckstrin homology domain within the endocytic dynamin tetramers⁴⁰ (Extended Data Fig. 6f-h), and is a determinant of conformational heterogeneity for these and other dynamin-family proteins^{38,40}. The presence of disease alleles close to this interface suggests that these mutations may compromise receptor binding and that defects in the recruitment of DRP1 to mitochondria may contribute to pathogenesis. Accordingly, we found that the disease-associated G362D mutant of DRP1 (Fig. 2b) failed to co-assemble with MID49 and displayed altered assembly and conformational properties (Extended Data Figs. 6e, 8a, i-l).

Receptor interface 4 includes S611 (equivalent to S637 of isoform 1, Extended Data Fig. 7c), an intensively studied phosphorylation site for protein kinase A or Ca²⁺/calmodulin-dependent protein kinase $I\alpha^{45-47}$. Phosphorylation at this site, as well as phosphomimetic alleles of DRP1, have been reported to inhibit mitochondrial fission, although the mechanism of inhibition was unclear 45,48 . Our structure suggests that phosphorylation should inhibit MID49 binding, and accordingly we found that the phosphomimetic mutant, S611D, failed to co-assemble with MID49(126–454) under the same solution conditions that bound the wild-type protein (Fig. 2g–i).

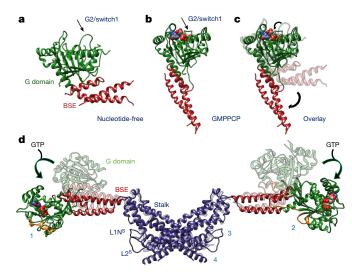


Fig. 3 | Nucleotide-driven allosteric elongation of DRP1 exposes MID49 and MID51 receptor binding sites. a, Nucleotide-free state of the DRP1 G domain and BSE as seen in a crystal structure (PDB ID: 4BEJ). The arrow points to the G2/switch1 loop. b, Conformation of GMPPCP-bound G domain and BSE determined by cryo-EM. c, Overlay of A and B. The curved arrows highlight the closing of the G2/switch 1 'lid' and the opening of the BSE 'wrist'. For comparison, the G2/switch 1 loop from 4BEJ was chosen from the only chain (B) in which it was completely resolved. d, Global conformational change induced by nucleotide binding. Rotation and translation of the G domain and BSE elongates the dimer and exposes receptor interfaces 1 and 2 (annotated on separate monomers for clarity). The surfaces of the G domains that engage the receptors are rendered orange in the nucleotide-bound and elongated conformation.

GTP-binding enables receptor binding

Understanding the allosteric coupling between the binding, hydrolysis and exchange of nucleotides and the conformational repertoire of dynamin-family GTPases remains a challenge. From the cryo-EM density we observed that the GMPPCP-bound G domains and the BSE of DRP1 adopt markedly different conformations compared to the nucleotide-free crystal structure²⁵. In addition to other nucleotide-induced conformational changes within the G domain, the most evident are the closing of the G2/switch-1 loop to form a closed 'lid' over the nucleotide (Fig. 3a,c). Analogous to the conformational change reported for dynamin-1⁴⁹, the closure of the switch-1 lid propagates through the adjacent $\beta\text{--sheet}$ to push the $\alpha\text{--helices}$ of the BSE into an orthogonal position (Supplementary Video 1). When evaluated in the context of the dynamin interface-2-containing X-shaped DRP1 dimer, this conformational change is a 90° rotation of the G domain and a 40 Å translation towards the stalk (Fig. 3d, Supplementary Video 2). Two of the four DRP1 surfaces that engage the DRR of MID49 or MID51 (receptor interfaces 1 and 2, Fig. 2a) are inaccessible in the nucleotide-free state but become available for binding upon nucleotidedriven elongation (Fig. 3d, Supplementary Video 2).

GTP hydrolysis induces filaments to curl into rings

We next evaluated the dynamics of the DRP1 + MID49 filaments in the presence of hydrolysable GTP, rather than the non-hydrolysable analogue used for 3D reconstruction. After copolymerization in the presence of the non-hydrolysable analogue, we exchanged GMPPCP for GTP by dialysis and followed the reaction using negative-stain transmission electron microscopy (TEM) at sequential time points until the GTP was consumed. We observed that the linear, three-sided DRP1–MID49 cofilaments disassembled into shorter, single-sided filaments before disassembling entirely upon complete hydrolysis to GDP (Fig. 4). The thinner single-sided filaments seen at intermediate time points resembled the single-sided filaments that we observed after mutagenizing a salt bridge that appears to hold the triangular structure together (Extended Data Fig. 3e–i). Moreover, the dynamics of

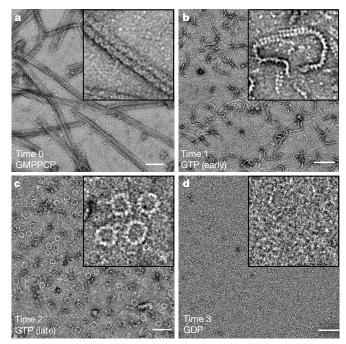


Fig. 4 | Dynamic instability of the DRP1–MID49 linear assembly and curling into closed DRP1 rings. a, Three-sided DRP1–MID49(126–454) linear filaments copolymerized with GMPPCP, as in Extended Data Figs. 2, 3. b, c, Subsequent exchange for GTP leads to disassembly of the three-sided filaments and partial disassembly of the single-sided filaments (b), and curling of single-sided filaments into closed rings (c). d, Complete consumption of GTP leads to complete oligomer disassembly. Scale bars, 100 nm

the single-sided filaments at intermediate time points were of interest (Fig. 4b, c). Specifically, upon reaching a reproducibly narrow range of lengths, the nearly linear single-sided filaments spontaneously curled into closed rings of markedly uniform diameter (Fig. 4c).

A model for closed DRP1 rings

In a separate but related experiment, we also evaluated the assembly properties of the DRP1 mutant G362D, which disrupts receptor interface 3, with and without MID49(126-454). As described above, this disease-associated residue sits at the base of the L1NS loop and this loop is a key site of inter-stalk interaction between adjacent DRP1 molecules in the linear filament (Figs. 1c, 2b, Extended Data Fig. 6b, e). We found that DRP1(G362D) purified as a nearly monodisperse and stable dimer, rather than as a mix of tetramers and higher-order species as observed for the wild-type, full-length protein (Extended Data Fig. 8a). In addition, DRP1(G362D) exclusively formed rings, not filaments, with or without MID49(126-454) and in the presence of GTP or GMPPCP (Fig. 5a, Extended Data Fig. 8i-l). These rings resembled those observed with wild-type DRP1 in all respects except that the wildtype protein formed closed rings via the linear MID49 copolymer only through the path-dependent reaction described above (Fig. 4c compared with Fig. 5a). We also observed that these apparently closed DRP1(G362D) rings could constrict liposomes into membrane tubules and circumscribe lipid nanotubes (Extended Data Fig. 9).

DRP1(G362D) rings showed improved structural homogeneity when formed using GMPPCP, presumably because when assembled with GTP, the rings remain dynamic and eventually disassemble upon hydrolysis to GDP (Fig. 4). We imaged the GMPPCP-bound DRP1(G362D) rings using cryo-EM and used 2D class averages of the predominant 12-dimer closed ring to model the differences between the linear filaments and the closed rings (Fig. 5, Extended Data Fig. 10). To account for the projected ring density, the G domain and the BSE of DRP1 must move even further down towards the stalks. Stalk interface-2 appears to remain constant in conformation, as revealed by the

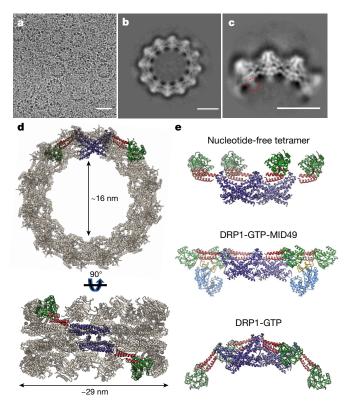


Fig. 5 | Drp1(G362D) cannot bind MID49 and forms rings exclusively with GMPPCP or GTP. a, Cryo-EM micrograph of Drp1(G362D) rings. b, 2D class average of the predominant closed ring that comprises 12 DRP1 dimers. c, 2D class average of a quarter of the ring revealing the secondary structure elements of the X-shaped DRP1 dimer. A red dashed circle indicates density that may be attributable to the variable domain. **d**, 3D model of the closed ring. **e**, Comparison between DRP1 tetramers observed in the nucleotide-free state (top, PDB: 4BEJ), the GMPPCP- and MID49(126-454)-bound linear state (middle), and the bent conformation modelled for the rings. Scale bars, 30 nm (a); 100 Å (b, c).

X-shaped dimer seen in projection (Fig. 5d, e). The curvature of the ring, however, dictates that stalk interfaces 1 and 3, and the conformations of the L1N^S and L2^S loops, must be extensively remodelled to allow an inter-dimer bending of approximately 30° in comparison with the linear DRP1-MID49 copolymer (Fig. 5d, e, Extended Data Fig. 10e, f, Supplementary Videos 2, 3). We did not observe any density for MID49 in the wild-type rings that form by curling of the DRP1-MID49 cofilament in the presence of GTP, nor in our higher-resolution analysis of the DRP1(G362D) rings that form with or without MID49 present (Fig. 5b, c, Extended Data Figs. 8i-l, 10). This suggests that MID49 binding is incompatible with the curled state of the ring-shaped oligomer, and that constriction therefore requires receptor dissociation (Supplementary Video 3).

Discussion

We note that with an inner diameter of around 16 nm, the closed ring may be sufficient to sever a double-membrane mitochondrion if both the outer and inner membranes are compressed together. Alternatively, if inner membrane fission is distinct and precedes outer membrane fission, a 16-nm diameter suggests that a single membrane tubule would be stabilized by these rings. The structures we observe in vitro may therefore correspond to a highly constricted but pre-fission state observed in vitro⁵⁰ and in living cells when another dynamin-family protein, dynamin-2, is depleted²⁹. Constriction by DRP1, therefore, may stabilize the high degree of membrane curvature that is suitable for the recruitment and final fission event catalysed by additional dynaminfamily enzymes.

Together, these findings establish four advances. First, our cryo-EM structure and mutagenesis studies revealed how receptor proteins such as MID49 and MID51 recruit and stabilize a specific nucleotide-bound conformation of DRP1 and initiate polymerization of a cofilament. We speculate that the nearly linear properties of this polymer have adapted to encircle low-curvature mitochondrial tubules. Second, analysis of the DRP1-MID49 copolymer revealed how the binding of a guanine nucleotide induces a conformational rearrangement to expose an avid network of receptor-binding sites. We now understand these nucleotidedriven allosteric transformations in the context of both full-length and oligomeric DRP1. Third, a path-dependent constriction reaction revealed GTP-dependent conformational dynamics. In this reaction, nucleotide exchange and hydrolysis led to MID49/51 receptor dissociation, disassembly from the ends of the linear filament, and concomitant curling of the shortening filaments into closed rings. The requirement for MID49 or MID51 receptor dissociation before constriction may explain how overexpression of the MID receptors inhibits mitochondrial fission²¹. Fourth, analysis of a disease mutant in the L1N^S loop, DRP1(G362D), highlights this loop as a fundamental determinant of receptor binding as well as the inter-stalk interactions that govern oligomer geometry. Together, these observations reveal how DRP1 performs mechanical work by curling from linear filaments into closed rings around mitochondria.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0211-2.

Received: 31 July 2017; Accepted: 23 April 2018; Published online 13 June 2018.

- Mishra, P. & Chan, D. C. Mitochondrial dynamics and inheritance during cell division, development and disease. Nat. Rev. Mol. Cell Biol. 15, 634-646
- Bleazard, W. et al. The dynamin-related GTPase Dnm1 regulates mitochondrial fission in yeast. Nat. Cell Biol. 1, 298-304 (1999).
- Toyama, É. Q. et al. AMP-activated protein kinase mediates mitochondrial
- fission in response to energy stress. *Science* **351**, 275–281 (2016). Roy, M., Reddy, P. H., lijima, M. & Sesaki, H. Mitochondrial division and fusion in metabolism. Curr. Opin. Cell Biol. 33, 111-118 (2015).
- Twig, G. et al. Fission and selective fusion govern mitochondrial segregation and elimination by autophagy. EMBO J. 27, 433-446 (2008).
- Mao, K., Wang, K., Liu, X. & Klionsky, D. J. The scaffold protein Atg11 recruits fission machinery to drive selective mitochondria degradation by autophagy. Dev. Cell 26, 9-18 (2013).
- Chan, D. C. Fusion and fission: interlinked processes critical for mitochondrial health. Annu. Rev. Genet. 46, 265-287 (2012).
- van der Bliek, A. M., Shen, Q. & Kawajiri, S. Mechanisms of mitochondrial fission and fusion. Cold Spring Harb. Perspect. Biol. 5, a011072 (2013).
- Frank, S. et al. The role of dynamin-related protein 1, a mediator of mitochondrial fission, in apoptosis. Dev. Cell 1, 515–525 (2001).
- Ishihara, N. et al. Mitochondrial fission factor Drp1 is essential for embryonic development and synapse formation in mice. Nat. Cell Biol. 11, 958-966
- 11. Wakabayashi, J. et al. The dynamin-related GTPase Drp1 is required for
- embryonic and brain development in mice. J. Cell Biol. 186, 805-816 (2009). Chatel-Chaix, L. et al. Dengue virus perturbs mitochondrial morphodynamics to dampen innate immune responses. Cell Host Microbe 20, 342-356 (2016).
- 13. Kim, S. J. et al. Hepatitis B virus disrupts mitochondrial dynamics: induces fission and mitophagy to attenuate apoptosis. PLoS Pathog. 9, e1003722 (2013).
- Friedman, J. R. et al. ER tubules mark sites of mitochondrial division. Science 334, 358-362 (2011).
- Murley, A. et al. ER-associated mitochondrial division links the distribution of mitochondria and mitochondrial DNA in yeast. eLife 2, e00422 (2013).
- Lewis, S. C., Uchiyama, L. F. & Nunnari, J. ER-mitochondria contacts couple mtDNA synthesis with mitochondrial division in human cells. Science 353, aaf5549 (2016).
- Osman, C., Noriega, T. R., Okreglak, V., Fung, J. C. & Walter, P. Integrity of the yeast mitochondrial genome, but not its distribution and inheritance, relies on mitochondrial fission and fusion. Proc. Natl Acad. Sci. USA 112, E947-E956
- Labbé, K., Murley, A. & Nunnari, J. Determinants and functions of mitochondrial behavior. Annu. Rev. Cell Dev. Biol. 30, 357–391 (2014).
- Kraus, F. & Ryan, M. T. The constriction and scission machineries involved in mitochondrial fission. J. Cell Sci. 130, 2953-2960 (2017).
- Osellame, L. D. et al. Cooperative and independent roles of the Drp1 adaptors Mff, MiD49 and MiD51 in mitochondrial fission. J. Cell Sci. 129, 2170–2181 (2016).

- Palmer, C. S. et al. Adaptor proteins MiD49 and MiD51 can act independently of Mff and Fis1 in Drp1 recruitment and are specific for mitochondrial fission. *J. Biol. Chem.* 288, 27584–27593 (2013).
- Palmer, C. S. et al. MiD49 and MiD51, new components of the mitochondrial fission machinery. EMBO Rep. 12, 565–573 (2011).
- Koirala, S. et al. Interchangeable adaptors regulate mitochondrial dynamin assembly for membrane scission. *Proc. Natl Acad. Sci. USA* 110, E1342–E1351 (2013).
- Gandre-Babbe, S. & van der Bliek, A. M. The novel tail-anchored membrane protein Mff controls mitochondrial and peroxisomal fission in mammalian cells. Mol. Biol. Cell 19, 2402–2412 (2008).
- Fröhlich, C. et al. Structural insights into oligomerization and mitochondrial remodelling of dynamin 1-like protein. EMBO J. 32, 1280–1292 (2013).
- Mears, J. A. et al. Conformational changes in Dnm1 support a contractile mechanism for mitochondrial fission. *Nat. Struct. Mol. Biol.* 18, 20–26 (2011).
- Ingerman, E. et al. Dnm1 forms spirals that are structurally tailored to fit mitochondria. J. Cell Biol. 170, 1021–1027 (2005).
- Daumke, O. & Praefcke, G. J. Mechanisms of GTP hydrolysis and conformational transitions in the dynamin superfamily. *Biopolymers* 105, 580–593 (2016).
- Lee, J. E., Westrate, L. M., Wu, H., Page, C. & Voeltz, G. K. Multiple dynamin family members collaborate to drive mitochondrial division. *Nature* 540, 139–143 (2016).
- Losón, O. C., Song, Z., Chen, H. & Chan, D. C. Fis1, Mff, MiD49, and MiD51 mediate Drp1 recruitment in mitochondrial fission. *Mol. Biol. Cell* 24, 659–667 (2013).
- 31. Lackner, L. L., Horner, J. S. & Nunnari, J. Mechanistic analysis of a dynamin effector. *Science* **325**, 874–877 (2009).
- Faelber, K. et al. Crystal structure of nucleotide-free dynamin. Nature 477, 556–560 (2011).
- Ford, M. G., Jenni, S. & Nunnari, J. The crystal structure of dynamin. Nature 477, 561–566 (2011).
- Richter, V. et al. Structural and functional analysis of MiD51, a dynamin receptor required for mitochondrial fission. J. Cell Biol. 204, 477–486 (2014).
- Losón, O. C. et al. The mitochondrial fission receptor MiD51 requires ADP as a cofactor. Structure 22, 367–377 (2014).
- Liu, R. & Chan, D. C. The mitochondrial fission receptor Mff selectively recruits oligomerized Drp1. Mol. Biol. Cell 26, 4466–4477 (2015).
- Losón, O. C. et al. Crystal structure and functional analysis of MiD49, a receptor for the mitochondrial fission protein Drp1. *Protein Sci.* 24, 386–394 (2015)
- Gao, S. et al. Structural basis of oligomerization in the stalk region of dynamin-like MxA. Nature 465, 502–506 (2010).
- Haller, O., Gao, S., von der Malsburg, A., Daumke, O. & Kochs, G. Dynamin-like MxA GTPase: structural insights into oligomerization and implications for antiviral activity. J. Biol. Chem. 285, 28419–28424 (2010).
- Reubold, T. F. et al. Crystal structure of the dynamin tetramer. Nature 525, 404–408 (2015).
- Chappie, J. S., Acharya, S., Leonard, M., Schmid, S. L. & Dyda, F. G domain dimerization controls dynamin's assembly-stimulated GTPase activity. *Nature* 465, 435–440 (2010).
- 42. Vanstone, J. R. et al. DNM1L-related mitochondrial fission defect presenting as refractory epilepsy. *Eur. J. Hum. Genet.* **24**, 1084–1088 (2016).
- Sheffer, R. et al. Postnatal microcephaly and pain insensitivity due to a de novo heterozygous DNM1L mutation causing impaired mitochondrial fission and function. Am. J. Med. Genet. A. 170, 1603–1607 (2016).

- Chang, C. R. et al. A lethal de novo mutation in the middle domain of the dynamin-related GTPase Drp1 impairs higher order assembly and mitochondrial division. *J. Biol. Chem.* 285, 32494–32503 (2010).
- Chang, C. R. & Blackstone, C. Dynamic regulation of mitochondrial fission through modification of the dynamin-related protein Drp1. Ann. NY Acad. Sci. 1201, 34–39 (2010).
- Chang, C. R. & Blackstone, C. Cyclic AMP-dependent protein kinase phosphorylation of Drp1 regulates its GTPase activity and mitochondrial morphology. J. Biol. Chem. 282, 21583–21587 (2007).
- Cribbs, J. T. & Strack, S. Reversible phosphorylation of Drp1 by cyclic AMP-dependent protein kinase and calcineurin regulates mitochondrial fission and cell death. *EMBO Rep.* 8, 939–944 (2007).
- Cereghetti, G. M. et al. Dephosphorylation by calcineurin regulates translocation of Drp1 to mitochondria. Proc. Natl Acad. Sci. USA 105, 15803–15808 (2008)
- Chappie, J. S. et al. A pseudoatomic model of the dynamin polymer identifies a hydrolysis-dependent powerstroke. Cell 147, 209–222 (2011).
- Ugarte-Uribe, B., Prevost, C., Das, K. K., Bassereau, P. & Garcia-Saez, A. J. Drp1
 polymerization stabilizes curved tubular membranes similar to those of
 constricted mitochondria. J. Cell Sci. 132, jcs208603 (2019).

Acknowledgements We thank M. Braunfeld, C. Kennedy, D. Bulkley, and A. Myasnikov and the University of California San Francisco (UCSF) Center for Advanced Cryo-EM, which is supported in part from National Institutes of Health (NIH) grants \$100D020054 and \$1\$100D021741 and the Howard Hughes Medical Institute (HHMI). We thank the QB3 shared cluster and NIH grant \$1\$100D021596-01, J.-P. Armache, N. Talledge for microscopy advice, C. Greenberg for consulting on structural modelling, D. Winge for discussions and M. Gu for facilitating mass spectrometry of proteins. This work was further supported by a Faculty Scholar grant from the HHMI (A.F.), the Searle Scholars Program (A.F.), NIH grant 1DP2GM110772-01 (A.F.), NIH grants GM53466 and GM84970 (J.M.S.), the Sandler Family Foundation through the UCSF Program for Breakthrough Biomedical Research and the American Asthma Foundation, and the HHMI (R.Y.-R.W., J.M.S. and D.A.A.). R.Y.-R.W. is an HHMI Fellow of the Life Sciences Research Foundation. A.F. is a Chan Zuckerberg Biohub investigator.

Reviewer information *Nature* thanks M. Ryan, H. Zhou and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions R.K., J.M.S. and A.F. conceived the study. R.K., A.Y. and R.V.T. performed all experiments. R.K., R.Y.-R.W. and A.F. performed the computational analyses. D.A.A. advised R.Y.-R.W. and R.K. on model building. All authors evaluated the results and edited the manuscript. R.K. and A.F. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0211-2.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0211-2.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.F. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Construct design. Wild-type DRP1 isoform 2 sequence was purchased from DNASU (sequence ID HsCD00043627, UNIPROT identifier: O00429-3, also known as Dlp1a) and was cloned into pET16b plasmid (Novagen) between the Nde1 and BamH1 sites. The vector was kindly provided by the laboratory of W. Sundquist with a 10×-His tag followed by a PreScission protease site (Leu-Glu-Val-Leu-Phe-Gln-Gly-Pro). The wild-type MID49(126-454) sequence was amplified by PCR and cloned into a pGEX6p1 vector with an N-terminal glutathione-S-transferase tag followed by a PreScission protease site. Site-directed mutagenesis was performed on pET16b-DRP1 and pGEX6p1-MID49(126-454) using the Gibson cloning method to introduce mutations⁵¹. All constructs were verified using Sanger sequencing.

Protein purification. Protein purification was performed as described 52 . In brief, plasmids containing the wild-type DRP1 or MID49(126–454) sequence were transformed in the BL21-DE3 (RIPL) strain of *Escherichia coli*. The colonies were inoculated in LB culture medium and grown overnight. Secondary inoculations were performed the next morning in ZY medium for autoinduction 53,54 . The cultures were grown to an optical density at 600 nm (OD $_{600}$) of 0.8 at 37 °C in baffled flasks and were transferred to 19 °C to grow for another 12 h. The cultures were spun down and the bacterial pellets were used for protein purification immediately or stored at $-80\,^{\circ}\text{C}$.

Full-length DRP1 wild-type and mutant variants were purified as described previously for wild-type DRP1 with modifications²⁵. In brief, the bacterial pellets were resuspended in buffer A (50 mM HEPES/NaOH (pH 7.5), 400 mM NaCl, 5 mM MgCl₂, 40 mM imidazole, 1 mM dithiothreitol (DTT), 0.5 mg DNase (Roche) and protease inhibitors (10 mM pepstatin, 50 mM phenylmethylsulfonyl fluoride, 0.5 mM aprotinin and 2 mM leupeptin), followed by cell disruption with a probe sonicator. Lysates were cleared by centrifugation at 40,000g in Beckman JA 25.50 rotors for 60 min at 4 °C. The supernatant was filtered using a 0.45- μm filter and applied to Ni-NTA Agarose beads pre-equilibrated with buffer B (50 mM $\,$ HEPES/NaOH (pH 7.5), 400 mM NaCl, 5 mM MgCl₂, 40 mM imidazole, 1 mM DTT). Upon the application of the supernatant, the beads were washed with 20 column-volumes each of buffer B and buffer C (50 mM HEPES/NaOH (pH 7.5), 800 mM NaCl, 5 mM MgCl₂, 40 mM imidazole, 1 mM DTT, 1 mM ATP, 10 mM KCl) followed by buffer D (50 mM HEPES/NaOH (pH 7.5), 400 mM NaCl, 5 mM MgCl₂, 80 mM imidazole, 1 mM DTT, 0.5% (w/v) CHAPS detergent). A final pre-elution wash was performed with 20 column-volumes of buffer B. Bound DRP1 was eluted with buffer E (50 mM HEPES/NaOH (pH 7.5), 400 mM NaCl, 5 mM MgCl₂, 300 mM imidazole, 1 mM DTT) and dialysed overnight at 4°C against buffer B without imidazole in the presence of PreScission protease to cleave the N-terminal $10\times$ -His tag. The protein was re-applied to a Ni-NTA column pre-equilibrated with dialysis buffer and was observed to bind the column without the 10×-His tag as well. Subsequently, the protein was eluted with buffer B containing 80 mM imidazole.

Pure protein was concentrated with a 30-kDa molecular weight cut-off centrifugal concentration device (Millipore). In the final step, DRP1 was purified by size-exclusion chromatography on a Superdex-200 column (GE Healthcare) in buffer F containing 20 mM HEPES/ NaOH (pH 7.5), 300 mM NaCl, 2.5 mM MgCl₂ and 1 mM DTT. Fractions containing DRP1 were pooled, concentrated, flash-frozen as single-use aliquots in liquid nitrogen and stored at $-80\,^{\circ}\text{C}$. Exact masses for purified DRP1 proteins were validated by matrix-assisted laser desorption ionization-time of flight mass spectrometry.

MID49(126-454) was purified as described with the following modifications³⁷. pGEX6p1-MID49(126-454) plasmid DNA (human, UNIPROT identifier: isoform 1 Q96C03-1, also known as MIEF2) was transformed in BL21 (DE3) RIPL cells. The colonies were grown overnight in LB medium and secondary cultures were grown in ZY medium. Cells were grown to an OD_{600} of 0.8–1, collected by centrifugation and processed immediately or stored at $-80\,^{\circ}\text{C}$ as described above. The bacterial pellets were lysed as described above in MID-buffer A (50 mM Tris pH 8.0, 500 mM NaCl, 5% glycerol, 1 mM DTT and 0.1% (v/v) Triton X-100). The lysates were pre-cleared at 40,000g and filtered using a 0.45- μ m filter before applying to 3 ml glutathione Sepharose beads (GE Healthcare). After overnight binding to beads, the unbound protein was removed and the beads were washed using 20 column-volumes each of MID-buffer A and MID-buffer B (50 mM Tris pH 8.0, 500 mM NaCl, 5% glycerol, 1 mM DTT). The protein was eluted with MID-buffer C (50 mM Tris pH 8.0, 500 mM NaCl, 5% glycerol, 1 mM DTT and 20 mM reduced glutathione). The eluate was cleaved overnight with PreScission protease while dialysing against MID-buffer D (20 mM Tris pH 8.0, 100 mM NaCl, 5% glycerol, 1 mM DTT). Cleaved protein was further purified using ion-exchange chromatography using a Q Sepharose column (GE Healthcare). The low-salt buffer for ion exchange was the same as MID-buffer D and the high-salt buffer was MID-buffer E (20 mM Tris pH 8.0, 1 M NaCl, 5% glycerol, 1 mM DTT). The relevant MID49(126–454) fractions were pooled, concentrated and further purified using a size-exclusion chromatography column pre-equilibrated with MID-buffer F (20 mM Tris pH 8.0, 200 mM NaCl, 5% glycerol, 1 mM DTT). The fractions containing MID49(126–454) were pooled, concentrated, flash-frozen in liquid nitrogen and stored as single-use aliquots at $-80\,^{\circ}\mathrm{C}$.

Filament assembly, cryo-EM sample preparation, data acquisition and processing. To assemble DRP1-MID49(126-454) filaments, the proteins were mixed to a final concentration of $2\mu M$ each and maintained for an hour at room temperature. The mixture was dialysed against assembly buffer: 20 mM HEPES pH 7.5, 50 mM KCl, 3 mM MgCl₂, 1 mM DTT and 200 μ M β , γ -methyleneguanosine 5'-triphosphate sodium salt (GMPPCP) with or without 0.2% octyl-glucopyranoside (Anatrace). The filaments were observed using negative-stain TEM or cryo-EM after vitrification. Under these conditions, the mutant DRP1(S611D) failed to co-assemble with MID49, but upon further lowering the ionic strength to 25mM KCl, DRP1(S611D) displayed detectable but greatly reduced co-assembly compared to wild-type protein. For vitrification, the sample was applied to Quantifoil holey carbon grids (R2/2) using a Vitrobot Mark III with 3.5 μl of sample, a 5-s blotting time and a 0-mm offset at 19 °C and 100% humidity. Images were collected on an FEI T30 Polara operating at 300 kV at a magnification of 31,000×. Images were recorded on a Gatan K2 summit camera in super resolution mode for a final binned pixel size of 1.22 Å per pixel. The movies were dose-fractionated, contained 30-40 frames, had a total exposure time of 6-8 s with 0.2 s per frame and a per-frame dose of 1.1 to 1.4 electrons per Å². SerialEM was used to automate data collection 55 . The defocus range was $0.8-3\,\mu m$ under focus. The data was motion-corrected and dose-weighted using UCSF Motioncor2⁵⁶. Contrast transfer function (CTF) parameter estimation on the non-dose-weighted but motioncorrected stacks was carried out using CTFFIND4 and GCTF^{57,58}.

Filaments were boxed using the program e2helixboxer.py from the EMAN2 suite⁵⁹. Particle coordinates were used to extract discrete particles using RELION $1.3-1.4^{60}$ and all further processing was carried out within the RELION suite. Multiple rounds of 2D classification identified the well-ordered segments. 3D autorefine was run using a customized Relion1.2 version with the IHRSR algorithm implemented^{61,62}. The consensus helical structure was used to classify the particles without refining helical symmetry (using RELION 1.4), resulting in two major classes that differed slightly in rise and twist (Extended Data Fig. 2c). Particles from each class were selected and independently refined again with helical RELION 1.2 and IHRSR. Analysis of these reconstructions revealed that each structure comprised three linear filaments that bundle together to form a structure that resembled a triangle in cross-section (Extended Data Figs. 2, 3). The vertices of the triangle are formed through asymmetric interactions between the G domains in adjacent filaments. The triangular arrangement of the bundled helices is unlikely to correspond to a biologically meaningful architecture, and this structure cannot form if the MID49 receptor is embedded in the outer mitochondrial membrane.

To further improve the signal-to-noise ratio, each of the three filaments in each independent half-map was segmented, extracted, resampled on a common grid and summed using UCSF Chimera^{63–67}. The respective symmetrized but unfiltered half-maps from each class were again aligned to a common grid and summed according to the C2 symmetry axis of the DRP1 dimer. In a last step, relion_postprocess was used to add the resulting and fully symmetrized half-maps (Extended Data Fig. 2c). These half-maps and the final summed map, with differential B-factor sharpening per region (Extended Data Figs. 2c, 4, 5), were used for atomic modelling using Rosetta as described below.

For the projection structure of the DRP1(G362D) rings, 2μM protein was mixed in a 1:1 molar ratio with MID49(126-454) and was allowed to stand at room temperature for an hour. The mixture was dialysed against the assembly buffer (without detergent) overnight. The sample was collected after 12 h and vitrified using ultrathin 3-nm carbon support films (Ted Pella). For vitrification, a Mark III vitrobot was used with $3.5\,\mu l$ of sample, a 0-mm offset, 100% humidity and a 3.5-s blot time. The images were collected using an FEI TF20 microscope and SerialEM for automated data collection. The data were recorded with a Gatan K2 camera operating in super resolution mode to collect dose-fractionated movie stacks with a final binned pixel size of 1.234 Å per pixel. 40 frames were collected per stack (0.2 s and 1.42 electrons per $Å^2$ per frame). The movie stacks were motion-corrected and the parameters of the transfer function were estimated as described above. Approximately 2,000 particles were picked manually for initial 2D classification in RELION 1.4 and these averages were used as templates for further particle picking by Gautomatch (http://www.mrc-lmb.cam.ac.uk/kzhang/). Final 2D averages of the entire rings versus quarter segments of the rings were computed using Relion1.4. Liposome and nanotube reactions. Liposomes were made as described previously 52 . In brief, 1,2-dioleoyl-sn-glycero-3-phospho-L-serine (DOPS) was purchased from Avanti Polar Lipids. DOPS dissolved in chloroform was dried under a steady stream of nitrogen and dried under vacuum for an hour. The dried lipid film was resuspended in *n*-hexane and dried again under nitrogen. The

resulting lipid film was dried under vacuum for 4 h and was finally resuspended in 20 mM HEPES pH 7.5 and 150 mM KCl. The same protocol was followed for making nanotubes, in which the mixture contained 60% D-galactosyl- β -1,1'-N-nervonoyl-D-erythro-sphingosine (Galactosyl Ceramide), 30% DOPS and 10% Ni²⁺-NTA DOGS (1,2-dioleoyl-sn-glycero-3-[(N-(5-amino-1-carboxypentyl) iminodiacetic acid)succinyl] nickel salt).

For assembly reactions of DRP1(G362D) over lipid, 0.5–2 μM protein was incubated with liposomes or nanotubes for an hour and dialysed against the assembly buffer without detergent.

Model building. The general procedure for atomic model interpretation and validation using Rosetta were performed as described⁶⁸. To obtain an initial model for DRP1, the crystal structure of nucleotide-free DRP1 (PDB ID: 4BEJ)²⁵ was used for the stalk region and DRP1 G domain-BSE structures bound to GMPPCP (PDB ID: 3W6O) were used for the G-domain and BSE regions. Density-guided model completion for DRP1 was carried out with RosettaCM⁶⁹ using this hybridization of DRP1 crystal structures. A converged solution appeared from the low-energy ensemble of the complete models generated by RosettaCM. However, among the low-energy ensemble, residues 503–610 were found to be extremely flexible without cryo-EM density constraints and therefore were omitted for further coordinate refinement. For MID49, the highly homologous mouse MID49 crystal structure³⁷ (81.3% identity, PDB ID: 4WOY, Extended Data Fig. 5c, d) was used to generate a homology model using RosettaCM and used as the starting model.

To enable fragment-based, density-guided model refinement with missing residues (503–610, DRP1), Rosetta iterative local rebuilding tool was customized to disallow backbone rebuilding at breaks within a single chain. Multiple rounds of refinement were performed for each component against one half-map (training map), and the other half-map (validation map) was used to monitor overfitting according to the detailed procedure described in ref. 68 .

We further refined the model in the context of a full assembly that included eight identical copies of each protein, Mg^{2+} and nucleotide that included all possible inter-domain molecular interactions in the filament (Extended Data Fig. 5a, b). Pseudo-symmetry was used 70 to enable and facilitate the energy evaluations of all neighbouring interactions around the asymmetric unit (green model, shown in Extended Data Fig. 5a) for final model refinement of the full assembly. To this end, refinement was done against the training map. Finally, the half-maps were used to determine a weight for the density map that did not introduce overfitting. Using the weight and with the symmetry imposed, the whole assembly of DRP1 and MID49 was refined in the full map, followed by B-factors refinement 71 . Finally, quantification of the buried surface area and the number and nature of the bonds involved for each DRP1–MID49 interaction interface modelled by Rosetta were performed with the PISA server (http://www.ebi.ac.uk/pdbe/pisa/).

Visual evaluation of the model-to-map correspondence was carried out in UCSF Chimera using unfiltered and unsharpened maps, maps uniformly sharpened with a range of ad hoc B-factors, and maps processed with a model-based local sharpening and local low-pass filtering procedure to optimize contrast and the visibility of high-resolution features of the map⁷².

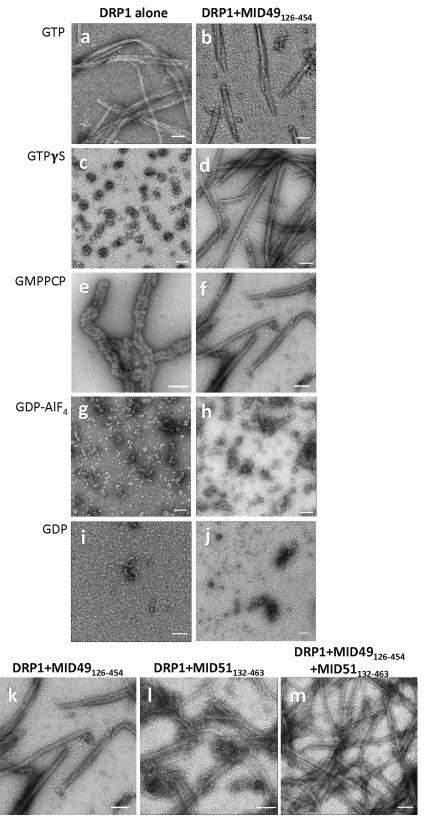
To build a molecular model for the closed 12-dimer DRP1 rings, we used the diameter, thickness and angles revealed by the 2D cryo-EM class averages of the DRP1(G362D) rings stabilized with GMPPCP. The atomic coordinates determined above using RosettaCM were used to build the ring in sections, first with repeating dimers of the interface-2 X-shaped stalk, then the BSE and finally the G domains and the angles between these sections were iteratively adjusted until calculated projections of the molecular model corresponded with the features of the experimental projection densities. Both the top (Fig. 5b, c) and the side view (Extended Data Fig. 10b) were used as constraints. The complete atomic model of the ring was finally refined in Phenix⁷³ to minimize clashes.

Statistics and reproducibility. All electron microscopy experiments in this this study were repeated at least three times.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

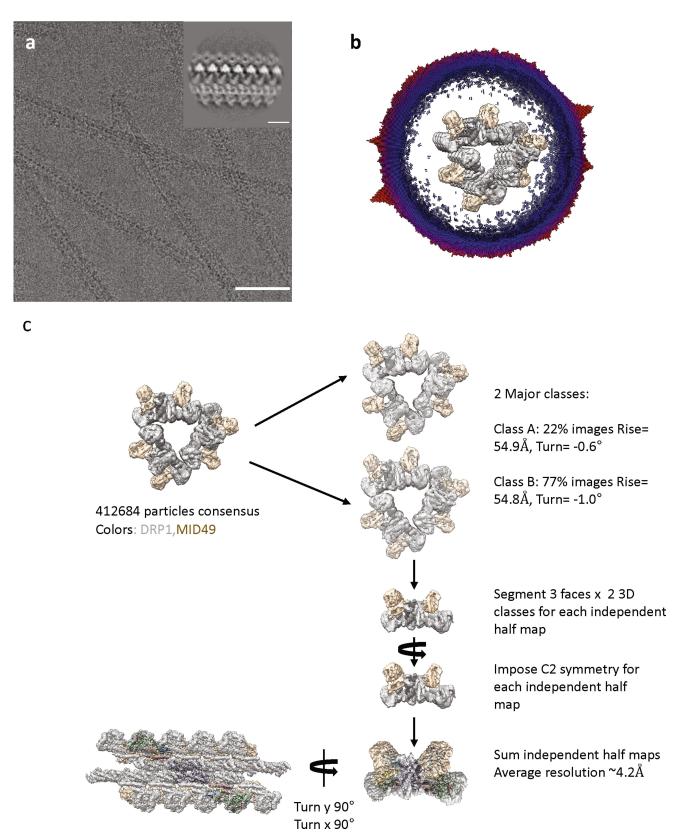
Data accessibility. All of the 3D cryo-EM density maps associated with this study have been deposited in the Electron Microscopy Data Bank with accession number EMD-8874. The atomic coordinates have been deposited in the Protein Data Bank as 5WP9. Raw data, models and image processing scripts are also available from the corresponding author upon reasonable request.

- 51. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
- Kalia, R., Talledge, N. & Frost, A. Structural and functional studies of membrane remodeling machines. Methods Cell Biol. 128, 165–200 (2015).
- Blommel, P. G., Becker, K. J., Duvnjak, P. & Fox, B. G. Enhanced bacterial protein expression during auto-induction obtained by alteration of lac repressor dosage and medium composition. *Biotechnol. Prog.* 23, 585–598 (2007).
- Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* 41, 207–234 (2005).
- 55. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
- Zheng, S. Q., Palovcak, E., Armache, J.-P., Cheng, Y. & Agard, D. A. MotionCor2: Anisotropic correction of beam-induced motion for improved single- particle electron cryo-microscopy. *Nat. Methods* 14, 331–332 (2017).
- Zhang, K. Gctf: Real-time CTF determination and correction. J. Struct. Biol. 193, 1–12 (2016).
- Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. J. Struct. Biol. 192, 216–221 (2015).
- Bell, J. M., Chen, M., Baldwin, P. R. & Ludtke, S. J. High resolution single particle refinement in EMAN2.1. Methods 100, 25–34 (2016).
- Scheres, S. H. W. Semi-automated selection of cryo-EM particles in RELION-1.3.
 J. Struct. Biol. 189, 114–122 (2015).
- Egelman, E. H. Reconstruction of helical filaments and tubes. *Methods Enzymol.* 482, 167–183 (2010).
- Ge, P. et al. Cryo-EM model of the bullet-shaped vesicular stomatitis virus. Science 327, 689–693 (2010).
- Pintilie, G. D., Zhang, J., Goddard, T. D., Chiu, W. & Gossard, D. C. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J. Struct. Biol.* 170, 427–438 (2010).
- Goddard, T. D., Huang, C. C. & Ferrin, T. E. Visualizing density maps with UCSF Chimera. J. Struct. Biol. 157, 281–287 (2007).
- Goddard, T. D., Huang, C. C. & Ferrin, T. E. Software extensions to UCSF chimera for interactive visualization of large molecular assemblies. Structure 13, 473–482 (2005).
- Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C. & Ferrin, T. E. Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics* 7, 339 (2006).
- Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612 (2004).
- Wang, R. Y. R. et al. Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. eLife 5, 1–22 (2016).
- Song, Y. et al. High-resolution comparative modeling with RosettaCM. Structure 21, 1735–1742 (2013).
- DiMaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling symmetric macromolecular structures in Rosetta3. PLoS ONE 6, e20450 (2011).
- DiMaio, F. et al. Atomic-accuracy models from 4.5-Â cryo-electron microscopy data with density-guided iterative local refinement. Nat. Methods 12, 361–365 (2015).
- Jakobi, A. J., Wilmanns, M. & Sachse, C. Model-based local density sharpening of cryo-EM maps. eLife 6, 1–26 (2017).
- Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D 66, 213–221 (2010).
- Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. Nat. Methods 11, 63–65 (2014).



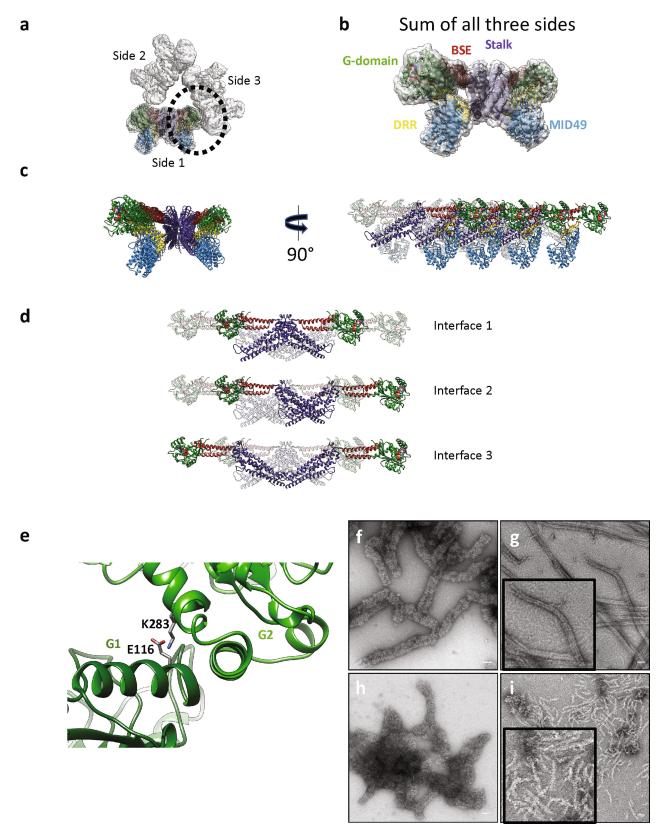
Extended Data Fig. 1 | DRP1 and MID49 assembly states. a–j, DRP1 assembly states visualized with negative-stain electron microscopy in the presence of different guanine nucleotides and MID49(126–454). Both proteins were incubated at concentrations of $2\,\mu M$. Scale bars, 100 nm.

 $\bf k-m$, MID49(126–454) and MID51(132–463) form indistinguishable assemblies with DRP1: DRP1 + MID49(126–454) and GMPPCP ($\bf k$), DRP1 + MID51(132–463) and GMPPCP ($\bf l$), DRP1 + both equimolar MID49 and MID51 ($\bf m$). Scale bars, 100 nm.



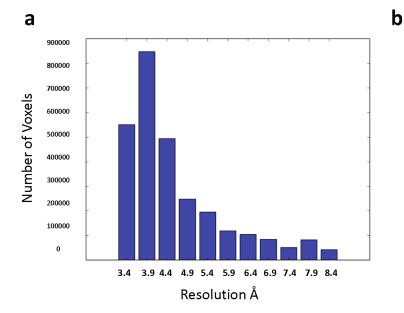
Extended Data Fig. 2 | Cryo-EM and 3D reconstruction. a, A cryo-EM micrograph of DRP1-MID49(126-454) filaments formed with GMPPCP. Scale bar, 100 nm. Inset, a representative 2D class average. Scale bar, 10 nm. b, Cross-section of the 3D reconstruction of the filament and the distribution of views determined during helical reconstruction. The length

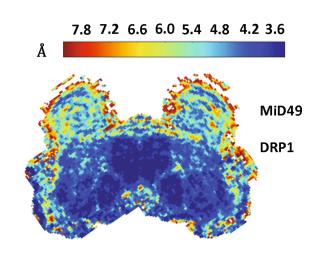
of the cylinders and the colour code correspond to the number of particles for that viewing direction (from few to many, blue to red). The 3D structure has been segmented and coloured with DRP1 in grey and MID49 in golden yellow. c, Particle numbers and workflow for the reconstruction protocol. DRP1 density is shown in grey and MID49 is in golden yellow.

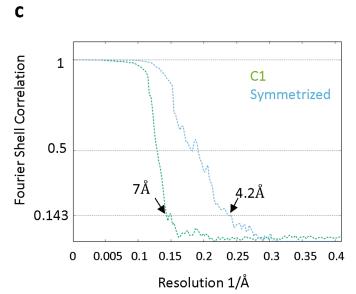


Extended Data Fig. 3 | Intra- and inter-filament interactions. a, The triangular structure seen in cross section. Side 1 of the triangular structure has the atomic model placed in the density. The G-domain-to-G-domain contact between adjacent sides is circled. b, The sum of the three sides with the model fit in density. c, Ribbon diagram of the same atomic model as in b. The rotated view shows eight chains each of DRP1 and MID49. The chains further from the reader are rendered transparent. d, An isolated tetramer of DRP1 from the filament, rendered to highlight the stalk interfaces 1, 2 and 3 observed for DRP1 and other GTPases of the

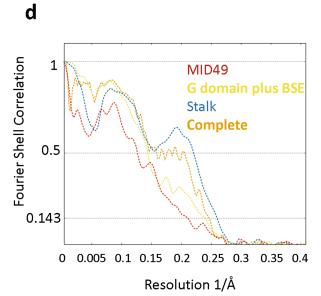
dynamin family. **e**, Expanded view of the circular region in **a** illustrating a salt bridge between adjacent G domains. **f-i**, Negative-stain micrographs of: DRP1-only wild-type polymers incubated with GMPPCP (**f**), DRP1 co-assembly with wild-type MID49(126–454) and GMPPCP (**g**; inset, high-magnification view), DRP1(E116R) mutant polymers (**h**), DRP1(E116R) mutant co-assembly with MiD49(126–454) (**i**). Shorter, single-sided filaments predominate. Disordered 'triangular assemblies' were also observed, but were much shorter and infrequent compared with the wild-type proteins. Scale bars, 50 nm.







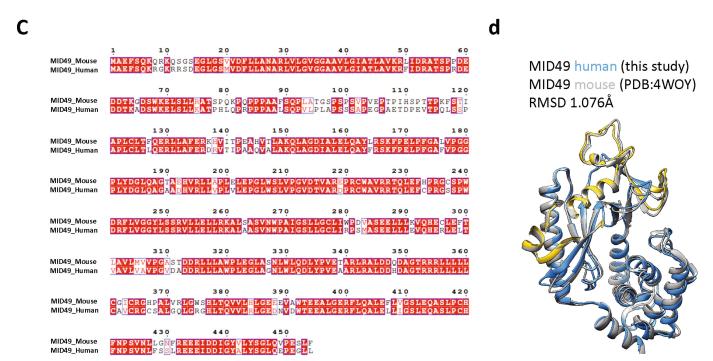
Extended Data Fig. 4 | **Resolution estimates. a**, **b**, Local resolution estimates computed by Resmap⁷⁴. Histogram of voxel values (**a**), and results in **a** depicted as a heat map of a cross-section through the



reconstruction (**b**). **c**, **d**, Fourier shell correlation plots for the half-maps with and without symmetry (**c**) and model-to-map correlations for each sub-region of the structure (**d**).

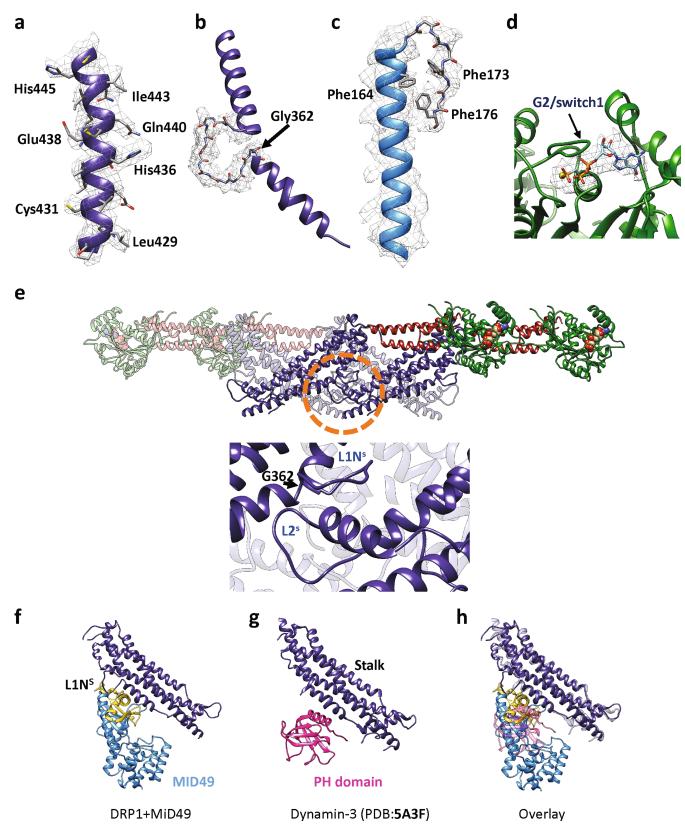
DRP1
MID49

DRP1
MID49



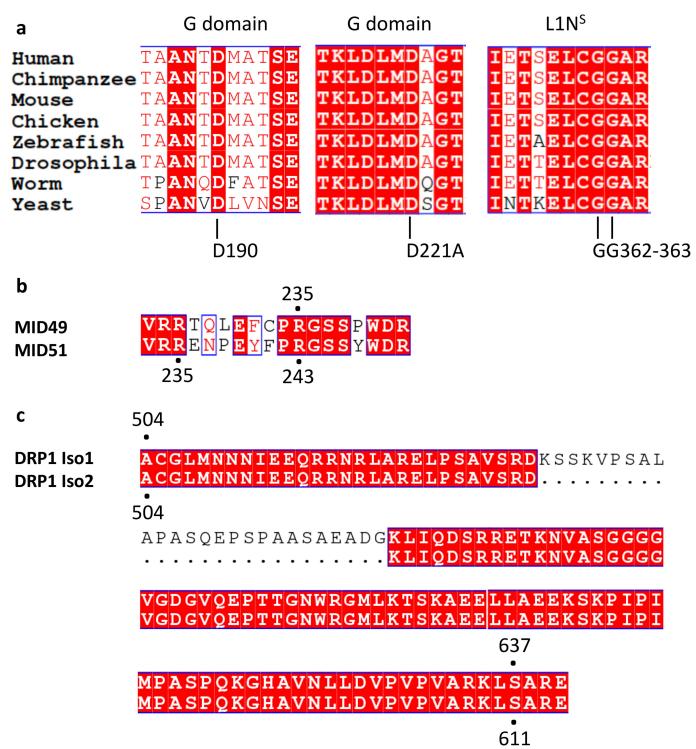
Extended Data Fig. 5 | Rosetta-based model refinement. a, Complete assembly used for Rosetta-based model building with the asymmetric unit shown in green. b, Atomic B-factors for one asymmetric unit, DRP1 versus MID49 models (ribbon) and bound GMPPCP (space filling). c, Sequence alignment between human and mouse MID49 sequences.

d, Overlay of the homology model of human MID49(126–454) (blue, with DRR in yellow, ribbon) modelled within the cryo-EM density overlaid with the mouse MID49 crystal structure (PDB: 4WOY, grey ribbon)³⁷ which was used as a constraint. No density attributable to ADP within the nucleotidyltransferase domain was observed.



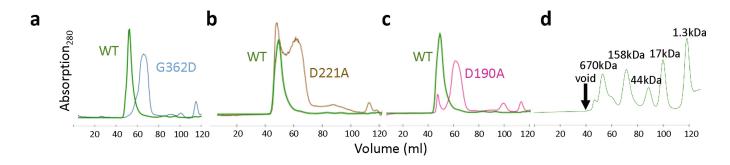
Extended Data Fig. 6 | Map-to-model fits and role of the L1N^S and L2^S loops. a–d, Examples of models fit within B-factor-sharpened cryo-EM density for a helix from the DRP1 stalk (a), the backbone of the L1N^S loop from the stalk (b), an elongated helix and turn found in MID49 (c) and GMPPCP and Mg within the G domain (d). e, Roles of L1N^S and L2^S in linear filament formation. Top, an isolated Drp1 tetramer from

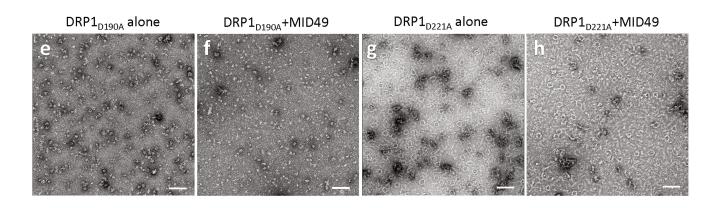
the cryo-EM model. The circled region is expanded in the lower panel. Bottom, interactions that the conserved loop L1N $^{\rm S}$ makes within the assembly. G362 is highlighted with an arrow. f, DRP1 stalk and MID49 at receptor interface-3. g, Dyn3 stalk and pleckstrin homology domain from PDB 5A3F 40 . h, Overlay of f and g.

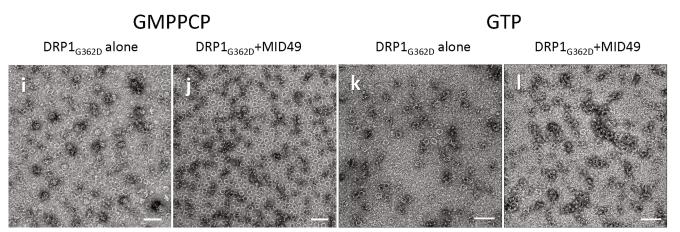


Extended Data Fig. 7 | **Sequence conservation and key interaction sites. a**, Multiple sequence alignment of the regions near and including the DRP1 residues mutated in this study: D190, D221 and G362, G363. The residue numbers apply to human DRP1, isoform 2 (UNIPROT identifier: O00429-3 which is also known as DLP1a). **b**, Sequence alignment of

MID49 and MID51 at the region around residue R235 of MID49. R235 of MID49 corresponds to R243 of MID51. **c**, Sequence alignment of DRP1 isoform 1 and isoform 2 showing the correspondence of S637 (isoform 1) and S611 (isoform 2).

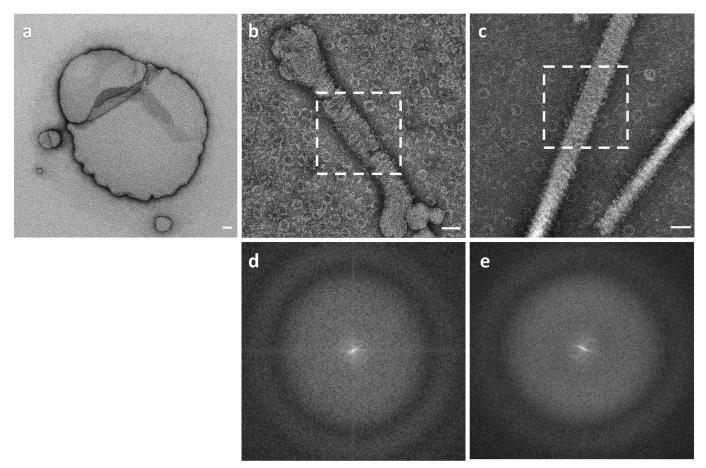






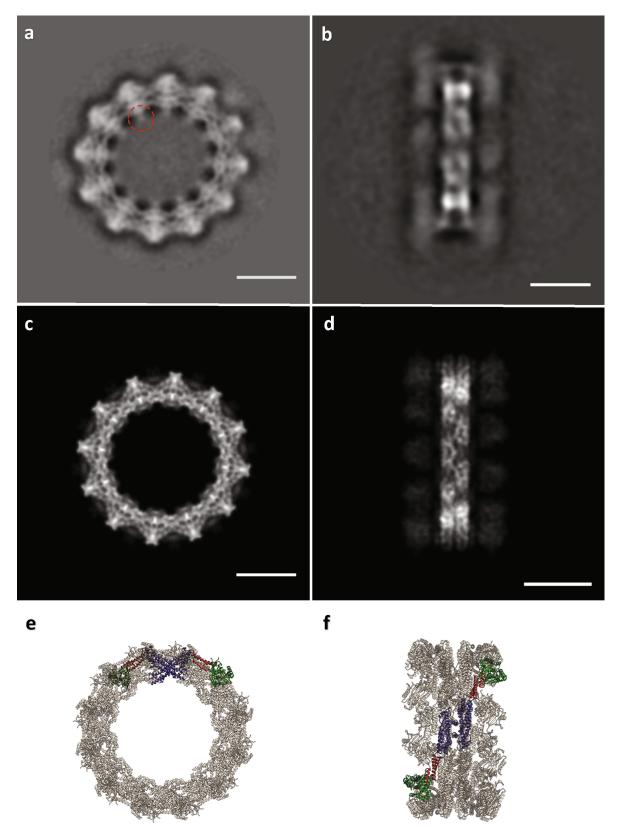
Extended Data Fig. 8 | Biochemical and structural characterization of mutants. a-c, Size-exclusion chromatography traces for DRP1 wild-type and mutants used in the study. Comparison between wild-type (WT) and G362D (a), wild-type and D221A (b), and wild-type and D190A (c). d, Gel filtration standards with annotated molecular weights from the same column and chromatography system. e-h, DRP1 assembly and co-assembly reactions with GMPPCP for DRP1(D190A) alone (e),

 $\begin{array}{l} DRP1(D190A) + MID49(126-454) \ (\textbf{f}), DRP1(D221A) \ alone \\ (\textbf{g}) \ and \ DRP1(D221A) + MID49(126-454) \ (\textbf{h}). \ Scale \ bars, 100 \ nm. \\ \textbf{i-l}, DRP1(G362D) \ assembly \ and \ co-assembly \ reactions \ with \ GMPPCP \ or \ GTP. \ DRP1(G362D) \ forms \ rings \ but \ not \ linear \ filaments \ without \ MID49 \\ (\textbf{i}, \textbf{k}), \ with \ MID49 \ (\textbf{j}, \textbf{l}), \ with \ GMPPCP \ (\textbf{i}, \textbf{j}) \ or \ with \ GTP \ (\textbf{k}, \textbf{l}). \ Scale \ bars, 100 \ nm. \end{array}$



Extended Data Fig. 9 | DRP1(G362D) rings on model membranes. a, DOPS liposomes used in the study. b, DOPS liposomes after incubation with DRP1(G362D) showing ring-like assemblies on the membrane and in the background. c Lipid nanotubes incubated with DRP1(G362D).

d, Power spectrum of the area shown within the dashed square in b.
e, Power spectrum of the area shown within the dashed square in c.
In both d and e, layer lines indicative of helical geometry are not detectable. Scale bars, 50 nm.



Extended Data Fig. 10 | DRP1(G362D) forms 12-dimer closed rings. a, 2D class average of the rings. The red dashed circle indicates density that may be attributable to the variable domain. b, 2D class average of infrequent, orthogonal or side views used as a constraint during model

building. c–f, Top (c) and side (d) projections of the model; top (e) and side (f) views of the final model rendered as ribbons. Scale bars, 100 Å. Green, G domain; red, BSE; purple, stalk.



Observations of the missing baryons in the warm-hot intergalactic medium

F. Nicastro^{1,2}*, J. Kaastra³, Y. Krongold⁴, S. Borgani^{5,6,7}, E. Branchini⁸, R. Cen⁹, M. Dadina¹⁰, C. W. Danforth¹¹, M. Elvis², F. Fiore¹, A. Gupta¹², S. Mathur¹³, D. Mayya¹⁴, F. Paerels¹⁵, L. Piro¹⁶, D. Rosa-Gonzalez¹⁴, J. Schaye¹⁷, J. M. Shull¹¹, J. Torres-Zafra¹⁸, N. Wijers¹⁷ & L. Zappacosta¹

It has been known for decades that the observed number of baryons in the local Universe falls about 30-40 per cent short^{1,2} of the total number of baryons predicted³ by Big Bang nucleosynthesis, as inferred^{4,5} from density fluctuations of the cosmic microwave background and seen during the first 2-3 billion years of the Universe in the so-called 'Lyman α forest'^{6,7} (a dense series of intervening H I Lyman α absorption lines in the optical spectra of background quasars). A theoretical solution to this paradox locates the missing baryons in the hot and tenuous filamentary gas between galaxies, known as the warm-hot intergalactic medium. However, it is difficult to detect them there because the largest by far constituent of this gas-hydrogen-is mostly ionized and therefore almost invisible in far-ultraviolet spectra with typical signal-to-noise ratios^{8,9}. Indeed, despite large observational efforts, only a few marginal claims of detection have been made so far^{2,10}. Here we report observations of two absorbers of highly ionized oxygen (O VII) in the high-signal-to-noise-ratio X-ray spectrum of a quasar at a redshift higher than 0.4. These absorbers show no variability over a two-year timescale and have no associated cold absorption, making the assumption that they originate from the quasar's intrinsic outflow or the host galaxy's interstellar medium implausible. The O vII systems lie in regions characterized by large (four times larger than average¹¹) galaxy overdensities and their number (down to the sensitivity threshold of our data) agrees well with numerical simulation predictions for the long-sought warmhot intergalactic medium. We conclude that the missing baryons have been found.

Numerical simulations in the framework of the commonly accepted (Λ CDM) cosmological paradigm predict that, starting at a redshift of $z\approx 2$ and during the continuous process of structure formation, diffuse baryons in the intergalactic medium (IGM) condense into a filamentary web (with electron densities of $n_e\approx 10^{-6}-10^{-4}~\rm cm^{-3}$) and undergo shocks that heat them up to temperatures of $T\approx 10^5-10^7~\rm K$, making the by-far-largest constituent of the IGM, hydrogen, mostly ionized^{8,9}. At the same time, galactic outflows powered by stellar and active galactic nucleus (AGN) feedback enrich the IGM baryons with metals¹². How far from galaxies these metals roam depends on the energetics of these winds, but it is expected that metals and galaxies are spatially correlated.

This shock-heated, metal-enriched medium, known as the warmhot intergalactic medium (WHIM), is made up of three observationally distinct phases: (1) a warm phase, with $T \approx 10^5 - 10^{5.7}$ K, where neutral hydrogen is still present with an ion fraction of $f_{\rm HI} > 10^{-6}$, and the best observable metal ion tracers are O vI (with main transitions in the far ultraviolet; FUV) and C v (with transitions in the soft X-rays); (2) a hot phase with $T \approx 10^{5.7} - 10^{6.3}$ K, where $f_{\rm HI} \approx 10^{-6} - 10^{-7}$, and O vII (with

transitions in the soft X-rays) largely dominates metals with ion fractions near unity; and (3) an even hotter phase ($T \approx 10^{6.3}$ – 10^7 K), coinciding with the outskirts of massive virialized groups and clusters of galaxies, where H I and H-like metals are present only in traces⁹. The warm phase of the WHIM has indeed been detected and studied in detail in the past few years and is estimated to contain an additional 15^{+8}_{-4} % fraction of the baryons (for example, see refs ^{1,2} and references therein; Table 1). This brings the total detected fraction to 61_{-12}^{+14} %, but still leaves us with a large $(39^{+12}_{-14}\%)$ fraction of elusive baryons, which if theory is correct—should be searched for in the hotter phases of the WHIM. In particular, the diffuse phase at $T \approx 10^{5.7}$ – $10^{6.3}$ K should contain the vast majority of the remaining WHIM baryons, and it is traced by O vII. Optimal signposts for this WHIM phase are then the O VII He α absorption lines; however, these are predicted to be relatively narrow (with a Doppler parameter thermal component $b_{\rm th}({\rm O}) \approx 20{\text -}46~{\rm km~s^{-1}}$), extremely shallow (rest-frame equivalent widths, EW $\lesssim 10 \text{ mÅ}$) and rare^{8,9}. Such lines are unresolved by current X-ray spectrometers and need a signal-to-noise ratio per resolution element greater than or equal to \sim 20 in the continuum to be detected at a single-line statistical significance (that is, before accounting for redshift trials; see Methods) exceeding $\sim 3\sigma$. This requires multi-million-second exposures against the brightest possible targets that are available at sufficiently high redshift ($z \gtrsim 0.3$).

Here we report on the longest observation performed with the X-ray multi-mirror mission (XMM)-Newton reflection grating spectrometer (RGS) on a single continuum target, the brightest X-ray blazar 1ES 1553+113, with z>0.4 (see Methods). The quality of this RGS spectrum makes it a goldmine for X-ray absorption-line studies (see Methods and Extended Data Figs. 1 and 2). We detect a number of unresolved absorption lines in both RGS units (RGS1 and RGS2) and at single-line statistical significance exceeding $2.7\,\sigma{-}3\sigma$. (Throughout the paper, we report ranges of statistical significance, where the upper boundary is the actual measured single-line statistical

Table 1 | Cosmic baryon census at z < 0.5

	$\Omega_{\rm b}h^2$	$\Omega_{\rm b}/\Omega_{\rm b}^{{\rm Plank}}$
Stars in galaxies	0.0015 ± 0.0004	(7±2)%
Cold gas in galaxies	0.00037 ± 0.00009	$(1.7 \pm 0.4)\%$
Galaxies' hot disks/haloes	0.0011 ± 0.0007	$(5\pm 3)\%$
Hot ICM	0.00088 ± 0.00033	$(4.0 \pm 1.5)\%$
Photoionized Lyman α forest	0.0062 ± 0.0024	$(28 \pm 11)\%$
WHIM with $10^5\text{K}\!\leq\!\textit{T}\!<\!10^{5.7}\text{K}$	$0.0033^{+0.0018}_{-0.0009}$	$15^{+8}_{-4}\%$
WHIM with $10^{5.7} \text{ K} \leq T < 10^{6.2} \text{ K}$	>0.002 and <0.009	>9% and <40%
Total	>0.013 and < 0.026	>59% and <118%

 $\Omega_{\rm b}$, baryon density; $\Omega_{\rm b}^{\rm Plank}$, total baryon density measured by the Planck satellite.

¹Istituto Nazionale di Astrofisica (INAF), Osservatorio Astronomico di Roma, Rome, Italy. ²Harvard–Smithsonian Center for Astrophysics, Cambridge, MA, USA. ³SRON Netherlands Institute for Space Research, Utrecht, The Netherlands. ⁴Instituto de Astronomia Universidad Nacional Autonoma de Mexico, Mexico City, Mexico. ⁵Physics Department, University of Trieste, Trieste, Italy. ⁵INAF—Osservatorio Astronomico di Trieste, Trieste, Italy. ³Physics Department, University of Roma Tre, Rome, Italy. ³Pepartment of Astrophysical Science, Princeton University, Princeton, NJ, USA. ¹¹OINAF—Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Bologna, Italy. ¹¹Department of Astrophysical Science, University of Colorado, Boulder, CO, USA. ¹²Columbus State Community College, Columbus, OH, USA. ¹³Ohio State University, Columbus, OH, USA. ¹⁴Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico. ¹⁵Department of Astronomy, Columbia University, New York, NY, USA. ¹⁵INAF - Istituto di Astrofisica e Planetologia Spaziali, Rome, Italy. ¹²Leiden Observatory, Leiden, The Netherlands. ¹³Instituto de Astrofisica de La Plata (IALP-UNLP), La Plata, Argentina. *e-mail: fabrizio.nicastro@oa-roma.inaf.it

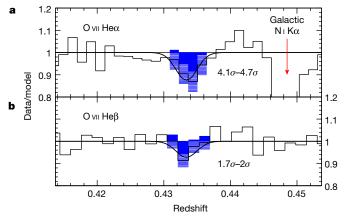


Fig. 1 | Intervening absorber at z_1^X = 0.4339 \pm 0.0008. a, b, Ratios of XMM-Newton RGS1 and RGS2 spectra of the blazar 1ES 1553+113 with their local best-fitting continuum model, showing the two O VII He α (a) and He β (b) absorption lines that identify System 1 at a 'true' statistical significance of 3.9σ -4.5 σ (these statistical significance boundaries correspond to Gaussian probabilities of chance detection of P= 4.8 \times 10⁻⁵ and P= 3.4 \times 10⁻⁶, respectively). Shaded blue regions indicate the lines detected in the X-ray region. Hatched blue intervals in the line histograms represent \pm 1 σ errors (statistical plus 2% systematic errors). Black curves are best-fitting Gaussians folded through the instrumental RGS line spread function.

significance, whereas the lower boundary is the measured significance, conservatively corrected for observed systematic errors in the RGS spectrum; see Methods and Extended Data Fig. 3). Particularly, two of these lines are seen at significances of $4.1\sigma-4.7\sigma$ (Figs. 1a) and $3.7\sigma-4.2\sigma$ (Fig. 2b) at wavelengths where no Galactic absorption is expected and no instrumental feature is present (Extended Data Fig. 2;

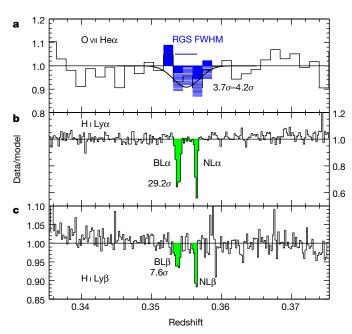


Fig. 2 | Intervening absorber at $\mathbf{z}_2^{\mathbf{X}} = \mathbf{0.3551}_{-0.0015}^{+0.0003}$. \mathbf{a} – \mathbf{c} , Same as Fig. 1, but showing the results for an O VII He α absorber (\mathbf{a}) and two H I Ly α (\mathbf{b}) and Ly β (\mathbf{c}) absorbers only \sim 750 km s⁻¹ apart and both at redshifts consistent with the X-ray System 2. The X-ray absorber has a 'true' statistical significance of 2.9σ – 3.7σ ; these statistical significance boundaries correspond to Gaussian probabilities of chance detection of $P=1.9\times10^{-3}$ and $P=10^{-4}$, respectively. Neither of the two H I absorbers in \mathbf{b} and \mathbf{c} can be physically associated to the O VII He α absorber (\mathbf{a}), implying the presence of at least three co-located—but dramatically different physically—gaseous phases.

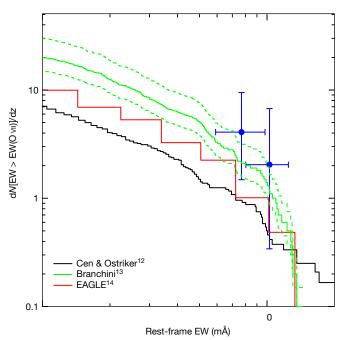


Fig. 3 | **Agreement between data and predictions.** Cumulative number density, N, of O VII He α intergalactic absorbers per unit redshift, z, with an EW greater than a given threshold, versus the rest-frame EW threshold. The lines show different predictions $^{12-14}$, compared with the two data points (blue points) that correspond to our two detections at $z_1^X = 0.3551$ and $z_2^X = 0.4339$. All error bars correspond to 1σ and vertical error bars are computed using low-number Poisson statistics. The black curve shows predictions from simulations performed with galactic super-winds and without imposing local thermodynamic equilibrium 12 . The green curve corresponds to a range of predictions from ref. 13 . The red line shows EAGLE (Evolution and Assembly of Galaxies and their Environments) results for the 100 Mpc reference simulation of ref. 14 at z = 0.1, and an O VII He α Doppler parameter b = 100 km s $^{-1}$ is used in the conversion from column density to EW.

see Methods). We attribute these lines to intervening O VII He α absorbers at redshifts of $z_1^X=0.4339\pm0.0008$ (hereafter, System 1) and $z_2^X=0.3551^{+0.0003}_{-0.0015}$ (System 2) (Extended Data Table 1). Their statistical significances decrease to $3.5\sigma-4\sigma$ and $2.9\sigma-3.7\sigma$, respectively, after accounting for the number of redshift trials (see Methods). Interestingly, a lower-significance $(1.7\sigma-2\sigma)$ absorption line can be modelled at the wavelength where the O VII He β line for System 1 is expected (Fig. 1b, Extended Data Figs. 1, 2), which increases the 'true' statistical significance of System 1 to $3.9\sigma-4.5\sigma$ (see Methods).

Given the proximity of our two systems with the upper limit $z\lesssim 0.48$ that we estimate for the redshift of our target (see Methods), we cannot rule out that these two systems are imprinted by material intrinsic to the blazar environment that outflows from this environment at speeds lower than 0.05c-0.12c, where c is the speed of light. However, a number of reasons make this scenario implausible (see Methods, Extended Data Fig. 4 and Extended Data Table 2 for details). The identification of System 1 and System 2 as genuine WHIM/circumgalactic medium (CGM) systems seems much more reasonable.

By modelling the X-ray data with our hybrid-ionization models (see Methods), we estimate the temperatures (T^X) and the oxygen ($N_{\rm O}^{\rm X}$) and equivalent H ($N_{\rm H}^{\rm X}$; modulo metallicity) column densities for System 1 and System 2. We obtain $T_1^{\rm X} = (6.8^{+9.6}_{-3.6}) \times 10^5 {\rm K}$, $N_{{\rm O},1}^{\rm X} = (7.8^{+3.9}_{-2.4}) \times 10^{15}$, $N_{{\rm H},1}^{\rm X} = (1.6^{+0.8}_{-0.5})(Z/Z_{\odot})^{-1} \times 10^{19} {\rm cm}^{-2}$ for System 1 and $T_2^{\rm X} = (5.4^{+9.0}_{-1.7}) \times 10^5 {\rm K}$, $N_{{\rm O},2}^{\rm X} = (4.4^{+2.4}_{-2.0}) \times 10^{15}$ and $N_{{\rm H},2}^{\rm X} = (0.9^{+0.5}_{-0.4})(Z/Z_{\odot})^{-1} \times 10^{19} {\rm cm}^{-2}$ for System 2, where Z is the metallicity and Z_{\odot} denotes the metallicity of the Sun.

For both systems, these quantities are in good agreement with predictions for the hot phase of the WHIM^{8,9,12}. Moreover, the number of

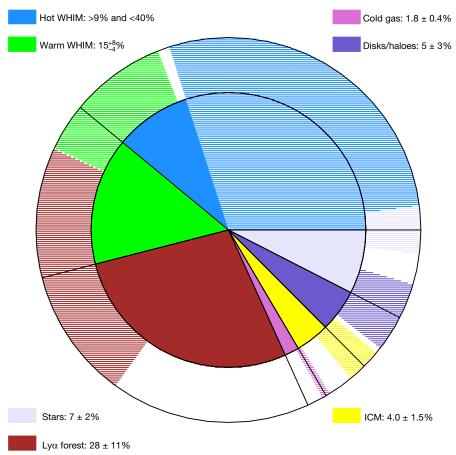


Fig. 4 | **Updated baryon census.** Pie diagram of the baryonic components of the local universe. Hatched regions in the external corona indicate the 90% uncertainties on the individual components; they are plotted across one of the two sides of each slice to show to what extent of each slice could be smaller or bigger. The exception is our measurement of the $10^{5.7}~{\rm K} \le T \le 10^{6.2}~{\rm K}$ WHIM component, where the solid-shaded

area shows the 3σ lower limit on the amount of baryons in these physical conditions, whereas the hatched area indicates and represents our 3σ upper limit, namely, the maximum amount of baryons that is allowed by our study to be contained in this phase at 99.97% confidence (see also Table 1).

detected O VII absorbers is consistent with predictions from a number of $models^{8,13,14}$ (Fig. 3).

The identification of the WHIM systems 1 and 2 is also supported by three independent pieces of evidence (see Methods for details): (1) significant galaxy overdensities at the locations of the two absorbers (Extended Data Figs. 5–7); (2) for System 1, the compatibility (at 2.3σ level) between the spectrum of 1ES 1553+113 obtained with the Hubble Space Telescope's cosmic origins spectrograph (HST-COS) and the presence of a very broad and shallow (and so physically consistent with O VII) H I Ly α absorber at the redshift of the X-ray absorber; (3) for System 2, the presence of two strong (but physically inconsistent with O VII) intervening hydrogen absorbers only 750 km s⁻¹ apart, and at redshifts consistent with that of the X-ray absorber (Fig. 2b, c).

Neither of the two strong H I absorbers seen in the HST-COS spectrum at redshifts consistent with that of System 2 (Fig. 2b) are sufficiently hot to produce the amount of O vII absorption seen in the X-rays (see Methods). This must be produced by even hotter gas (as indicated by X-ray data fitting), possibly confined between the two colder H I phases, that gives rise to undetectable broad H I Ly α absorption. For both X-ray absorbers, we use the HST-COS spectrum to derive 3σ upper limits on the column densities of H I and O vI (see Methods).

 3σ upper limits on the column densities of H I and O vI (see Methods). For System 1 we obtain $N_{\rm HI,1}^{\rm FUV} < 3.9 \times 10^{13} \, {\rm cm^{-2}}$ and $N_{\rm OVI,1}^{\rm FUV} < 3.2 \times 10^{13} \, {\rm cm^{-2}}$, and for System 2 we get $N_{\rm HI,2}^{\rm FUV} < 3.5 \times 10^{13} \, {\rm cm^{-2}}$ and $N_{\rm OVI,2}^{\rm FUV} < 8.1 \times 10^{13} {\rm cm^{-2}}$. Comparing $N_{\rm OVI}^{\rm FUV}$ with the 1σ lower boundary on $N_{\rm O}^{\rm X}$ allows us to constrain the minimum ionization correction needed (see Methods) and so to further limit the temperatures of the two systems in the intervals $T_{\rm I}^{\rm X} = (0.8-1.6) \times 10^6 \, {\rm K}$

and $T_2^{\rm X}=(0.5-1.4)\times 10^6{\rm K}$. Correcting the 3σ upper limits on $N_{\rm HI}^{\rm FUV}$ for the central values of the H I ionization fractions gives upper limits on the total H column densities of the two systems of $N_{\rm H,1}^{\rm FUV}<1.4\times 10^{20}~{\rm cm^{-2}}$ and $N_{\rm H,2}^{\rm FUV}<9\times 10^{19}~{\rm cm^{-2}}$, respectively. Finally, comparing these columns with those obtained from the X-ray data, $N_{\rm H}^{\rm X}$, we derive 3σ lower limits on the metallicity of the systems, $Z_1^{\rm X}>0.1Z_{\odot}$ and $Z_2^{\rm X}>0.1Z_{\odot}$ (Extended Data Table 3). For both systems we assume as upper limit on the average WHIM

For both systems we assume as upper limit on the average WHIM metallicity (see Methods) the value $Z_{\rm ICM}=0.2Z_{\odot}$ found in the peripheries (at r_{500} ; namely, the distance from the centre of the cluster where the density of the intra-cluster medium is 500 times the average density of the Universe.) of the intra-cluster medium 15 . With metallicity constrained in the $Z\approx0.1Z_{\odot}-0.2Z_{\odot}$ interval, we can now use the 68% confidence intervals on the equivalent H column densities to constrain the cosmological mass density of baryons with temperatures in the interval $T\approx10^{5.7}-10^{6.2}$ K. By parameterizing the lower limit by the inverse of the metallicity in units of $0.2Z_{\odot}$, we obtain a baryon density $\Omega_{\rm b}$ of $0.002(Z_{0.2})^{-1} < \Omega_{\rm b}^{10^{5.7}{\rm K} \le T \le 10^{6.2}{\rm K}} h^2 < 0.009$, corresponding to $9(Z_{0.2})^{-1}-40\%$ of the total baryon density measured by the Planck sattellite⁴, in good agreement with predictions^{8,9} and potentially sufficient to complete the baryon census (Fig. 4, Table 1).

Finally, theory predicts that metal-enriched WHIM absorbers should lie in the proximity of galaxy overdensities, either in the CGM of a particular galaxy or in the more diffuse IGM. Consistent with expectations, our photometric redshifts of the r'>23.5 galaxies in the $30'\times30'$ field surrounding 1ES 1553+113 indicate that both WHIM systems 1 and 2 are found in regions of substantial galaxy overdensities (Extended

Data Figs. 5–7; see Methods). For System 1 we have a number of spectroscopic redshift confirmations (that is, within $\pm 900~\rm km~s^{-1}$ of the absorber; see Methods) showing in particular the presence of a bright (Sloan Digital Sky Survey magnitude, i'=19.6) spiral at only 129 kpc and $-15~\rm km~s^{-1}$ from the absorber. For System 2, the only spectroscopically confirmed galaxy (out of only four spectroscopic redshifts available), a bright i'=20.5 elliptical, lies far (633 kpc and $+370~\rm km~s^{-1}$) from the absorber. This may explain the different baryon column densities of the two systems: System 1, with its higher column density, could be imprinted by the CGM of the nearby spiral, whereas System 2, with its lower baryon column density, could be produced by a more extended diffuse IGM filament connecting a structure of galaxies.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0204-1.

Received: 9 February 2018; Accepted: 25 April 2018; Published online 20 June 2018.

- Shull, J. M., Smith, B. D. & Danforth, C. W. The baryon census in a multiphase intergalactic medium: 30% of the baryons may still be missing. *Astrophys. J.* 759, 23 (2012).
- Nicastro, F., Krongold, Y., Mathur, S. & Elvis, M. A decade of warm hot intergalactic medium searches: where do we stand and where do we go? Astron. Nachr. 338, 281–286 (2017).
- Kirkman, D., Tytler, D., Suzuki, N., O'Meara, J. M. & Lubin, D. The cosmological baryon density from the deuterium-to-hydrogen ratio in QSO absorption systems: D/H toward Q1243+3047. Astrophys. J. Suppl. Ser. 149, 1–28 (2003).
- Planck Collaboration. Planck 2015 results. XIII. Cosmological parameters. Astron. Astrophys. 594, A13 (2016).
- Komatsu, E. et al. Five-year Wilkinson microwave anisotropy probe observations: cosmological interpretation. Astrophys. J. Suppl. Ser. 180, 330–376 (2009).
- Rauch, M. The Lyman alpha forest in the spectra of QSOs. Annu. Rev. Astron. Astrophys. 36, 267–316 (1998).
- Weinberg, D. H., Miralda-Escudé, J., Hernquist, L. & Katz, N. A lower bound on the cosmic baryon density. Astrophys. J. 490, 564–570 (1997).
- 8. Cen, R. & Ostriker, J. P. Where are the baryons? *Astrophys. J.* **514**, 1–6 (1999)
- Davé, R. et al. Baryons in the warm-hot intergalactic medium. Astrophys. J. 552, 473–483 (2001).
- Bonamente, M. et al. A possible Chandra and Hubble Space Telescope detection of extragalactic WHIM towards PG 1116+215. Mon. Not. R. Astron. Soc. 457, 4236-4247 (2016).

- Willmer, C. N. A. et al. The deep evolutionary exploratory probe 2 galaxy redshift survey: the galaxy luminosity function to z~1. Astrophys. J. 647, 853–873 (2006).
- Cen, R. & Ostriker, J. P. Where are the baryons? II. Feedback effects. Astrophys. J. 650, 560–572 (2006).
- Branchini, E. et al. Studying the warm hot intergalactic medium with gamma-ray bursts. Astrophys. J. 697, 328–344 (2009).
- Schaye, J. et al. The EAGLE project: simulating the evolution and assembly of galaxies and their environments. Mon. Not. R. Astron. Soc. 446, 521–554 (2015).
- Mernier, F. et al. Radial metal abundance profiles in the intra-cluster medium of cool-core galaxy clusters, groups, and ellipticals. Astron. Astrophys. 603, A80 (2017).

Acknowledgements This work is based on observations obtained with XMM-Newton, an ESA science mission with instruments and contributions directly funded by ESA Member States and NASA. F.N. and M.E. acknowledge support from NASA grant NNX17AD76G. S.M. acknowledges NASA grant NNX16AF49G. S.B. acknowledges financial support through agreement ASI-INAF n.2017-14-H.O and an INFN INDARK grant. Y.K. thanks INAOE for the support offered during a sabbatical visit in 2017 and acknowledges support from grant DGAPA-PAPIIT 106518 and from programme DGAPA-PASPA. R.C. acknowledges support from NSF grant AST-1515389.

Reviewer information *Nature* thanks R. Davé and T. Fang for their contribution to the peer review of this work.

Author contributions F.N. designed the study (together with J.K., L.P., S.B., L.Z., S.M., M.E., R.C., F.P. and A.G.), reduced and analysed the X-ray data, analysed the FUV data, extracted diagnostics by modelling the X-ray and FUV data and wrote the paper. Y.K., J.K., L.P., S.M., M.E., F.P., F.F. and A.G. helped with the analysis and modelling of the X-ray spectra. M.D. and F.N. designed and executed the Monte Carlo simulations used to evaluate the statistical significance of the absorbers. J.K. provided the most up-to-date XMM-Newton RGS calibrations and effective area corrections. C.W.D. and J.M.S. reduced the HST-COS data, extracted the G130 and G160 final spectra and helped with the interpretation of those spectra. Y.K. performed the optical photometric observations, as well as the reduction and analysis of those data, with the help of D.M. and D.R.-G. D.R.-G. and J.T.Z. provided the galaxies' spectroscopic redshifts. S.B., E.B., J.S., N.W. and R.C. provided hydrodynamical simulations and general theoretical support to the results. In particular, J.S. and N.W. provided results from highresolution Eagle simulations and F.P. helped to extract the number density of O VII absorbers from these simulations. All authors contributed equally to the discussion of the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0204-1

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to F.N. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Errors. Throughout the paper, uncertainties are quoted at 68% significance, unless stated otherwise.

Redshift of 1ES 1553+113. The exact redshift of this blazar is unknown, but a tight spectroscopic lower limit of z>0.413 is imposed by the detection up to this redshift (and close to the long-wavelength end of the HST-COS bandpass) of intervening H I Ly\$\alpha\$ absorption in the HST-COS spectrum of this target. Based on the lack of secure H I Ly\$\alpha\$ detections at z>0.413, the frequency of the H I Ly\$\alpha\$ absorbers (about one every 10 Å, down to the sensitivity of the regions of the HST-COS spectrum with the highest signal-to-noise ratio) and the decreasing signal-to-noise ratio of the HST-COS spectrum at \$\lambda>1,750 Å, we estimate a conservative upper limit of $z\lesssim0.48$. More recently, a $z=0.49\pm0.04$ redshift estimate has been set based on \$\gamma\$-ray observations of extragalactic background light absorption \$^{16}\$, consistent with our spectroscopic estimates.

Statistical significance. We define the 'single-line statistical significance' of an absorption line as the ratio EW/ Δ (EW) between the line equivalent width and its negative 1σ error (but see 'Additional systematic errors in the RGS spectrum of 1ES 1553+113' for a correction for effective-area systematic errors in the RGS spectrum). For a blind search of intervening absorbers, this is not the actual statistical significance of the line, unless a prior is used for the absorber redshift. In the absence of such a prior, we estimate the 'true' statistical significance of a given line by performing Monte Carlo simulations. Each simulation consists of: (a) producing 18 RGS1 and RGS2 mock spectra with the same exposures and continuum best-fitting parameters of our 18 XMM-Newton observations, (b) co-adding them and (c) fitting the co-added spectra (for RGS1 and RGS2) with a continuum model plus an unresolved and negative-only Gaussian, whose position is allowed to vary from the rest-frame position of the transition and its redshifted position at z = 0.5(this redshift is conservatively assumed to be the blazar redshift; also, for O VII He α lines, it coincides with the long-wavelength end of the RGSs). We repeat this procedure 10,000 times and evaluate the chance of detection of a line with single-line statistical significance greater than a given threshold.

We adopt this 'true' statistical significance for the X-ray line with the highest single-line statistical significance and use that line as a prior for other associated X-ray lines, when present. Finally, we evaluate the statistical significance of an X-ray absorption system by adding in quadrature the 'true' statistical significances of the lines of the system.

XMM-Newton RGS spectra of 1ES 1553+113. XMM-Newton observed 1ES 1553+113 for two consecutive cycles and for a total observing time of 1.75 Ms under a cycle-14 Very Large Program. Specifically, observations were made for 0.8 Ms between 29 July 2015 and 5 September 2015 (Epoch 1; black line in Extended Data Fig. 4a, b) and for 0.95 Ms between 1 February 2017 and 22 February 2017 (Epoch 2; red line in Extended Data Fig. 4a, b). Four short additional observations performed between 6 September 2001 and 28 July 2014 were also present in the archive, totalling an observing time of 0.096 Ms. We reduced these observations with the latest version (16.1.0) of the XMM-Newton Science Analysis System (SAS) software and current calibration files and extracted the RGS (RGS1 and RGS2) spectra and responses by applying the latest calibration corrections. We used the most up-to-date effective area corrections by switching on the parameters withrectification and witheffectiveareacorrection of the SAS tool rgsrgsrmfge that generates the RGS response matrices. The second of these parameters, in particular, eliminates some residual wiggle seen in the spectra of the calibration targets Mkn 421 and PKS 2155-304 at wavelength \sim 29-32 Å (Extended Data Fig. 2). We also added a 2% systematic uncertainty to the RGS counts, as estimated by the team of RGS calibrators, led by J.K. The final, cleaned RGS spectra have an exposure time of 1.85 Ms and are shown in Extended Data Fig. 1.

The RGS spectrum of 1ES 1553+113 has a signal-to-noise ratio per resolution element of SNRE = 33 at λ = 23.5–30.2 Å, and SNRE = 23 at λ = 21.6–23.5 Å (where only one of the two RGSs is present) and $\lambda = 30.2 - 32$ Å (the long-wavelength end of the RGS bandpass). This implies a 90% sensitivity of EW \geq 3.5 mÅ and EW \geq 4.5 mÅ to intervening O VII He $\!\alpha$ lines in the two redshift intervals $0.08 \le z \le 0.4$ and $z \in [(0, 0.08) \cup (0.4, 0.5)]$, respectively. The 8-33 Å RGS spectrum shows a number of narrow (unresolved) line-like negative features (Extended Data Fig. 1), eight of which are identifiable with Galactic absorption (marked and labelled in blue in Extended Data Fig. 1): Ne x Ly α (λ = 12.131 \pm 0.035 Å), Ne IX Heα (λ = 13.460 ± 0.035 Å), O vII Heα (λ = 21.586 ± 0.035 Å), O IV Kα + O I Kγ ($\lambda = 22.701 \pm 0.035\,$ Å), О 11 К α ($\lambda = 23.339 \pm 0.035\,$ Å), О 1 К α ($\lambda = 23.512 \pm 0.035$ Å), N vi K α ($\lambda = 28.772 \pm 0.035$ Å) and N i K α ($\lambda = 31.281 \pm 0.035$ Å), where half of the RGS full-width at half-maximum (FWHM) resolution element is used as position error. Two additional unresolved absorption lines are detected in both RGSs at combined single-line statistical significances of $4.1\sigma-4.7\sigma$ (Fig. 1a, Extended Data Figs. 1, 2 and Extended Data Table 1) and $3.7\sigma-4.2\sigma$ (Fig. 2a, Extended Data Figs. 1, 2 and Extended Data Table 1), at wavelengths where (1) no Galactic absorption is expected and (2) neither of the two spectrometers is affected by instrumental features due to cool

pixels in the dispersing detector (Extended Data Fig. 2). These are the lines here identified as intervening WHIM O VII He α at $z_1^{\rm X}=0.4339\pm0.0008$ (System 1) and $z_2^{\rm X}=0.3551^{+0.0003}_{-0.0015}$ (System 2; Extended Data Table 1 and Figs. 1, 2). An additional lower-significance (1.7 σ –2 σ) line is detected at a λ =26.69 ±0.09 Å and is identifiable as O VII He β at a redshift consistent with $z_1^{\rm X}$ (Fig. 1b, Extended Data Figs. 1, 2, and Extended Data Table 1).

Additional systematic errors in the RGS spectrum of 1ES 1553+113. In addition to the 2% systematic error estimated by the team of RGS calibrators for any RGS spectrum, we look for the presence of any systematic errors specific to our co-added 1.85-Ms RGS spectra of 1ES 1553+113, by performing the following Monte Carlo test. We re-fit the RGS spectra of 1ES 1553+113 by adding to the best-fitting continuum-plus-absorption-line models shown in Extended Data Fig. 1 an unresolved Gaussian (whose normalization is allowed to be either positive or negative) at a random position in the $\lambda = 8-33$ Å range. We then evaluate the single-line statistical significance of the line (assumed to be negative for emission lines and positive for absorption lines). There are about 400 RGS resolution elements in this wavelength range, and we therefore repeat this operation 1,000 times. Extended Data Fig. 3 shows the distribution of measured single-line significances (black histogram) and compares it with the expected distribution for a normal distribution with standard deviation of unity (red curve). The data distribution is symmetric (indicating that any systematic errors are also acting symmetrically) but slightly flatter than the red curve in Extended Data Fig. 3. We think that this is due to uncertainties in the RGS effective areas at wavelengths corresponding to cool pixels in the dispersing detectors. This effect can indeed be seen in Extended Data Fig. 2, where the residual excesses or deficits of counts in the data at, for example, $\lambda \approx 27.7$ Å, 30.2–3.3 Å and 31.1–31.2 Å, all correspond to strong effective-area instrumental features (red and black curves). The normal distribution that best fits our Monte Carlo results has a standard deviation of 1.15 (green curve in Extended Data Fig. 3), which should then be used to correct the statistical significance of lines detected at wavelengths affected by cool pixels. The absorption lines reported here are all found in relatively clean effective-area spectral regions. However, to be conservative, for each line we use a range of statistical significances, whose lower and upper boundaries are the measured single-line statistical significances corrected for effective-area-induced systematic errors and the actual single-line significance, respectively. Such single-line statistical significance intervals are then propagated to 'true' statistical significance intervals by allowing for redshift trials, as detailed above.

X-ray diagnostics: temperature and baryon column density. We use our hybrid-ionization models (that is, models of collisionally ionized gas perturbed by photoionization, at a given redshift, by the meta-galactic radiation field)¹⁷, to characterize the temperature T^{X} and the equivalent H column density N_{H}^{X} (modulo the absorber's absolute metallicity) of the X-ray absorbers. In our models we use relative metallicities from ref. $^{\rm 18}.$ Our spectral WHIM models include more than 3,000 line transitions in the soft X-ray band (energies $E \approx 0.1-2$ keV). For each transition, the line optical depth and Voigt profile are self-consistently computed for pairs of values of the gas temperature (that is, ionization structure) and equivalent H column density, which are let free to vary in the fit. At typical WHIM conditions, the temperature of these models is well constrained by the detected ion transitions, as well as by the upper limits that can be set on the presence of lower- or higher-ionization ion transitions. For example, in the 8–33 Å band covered by the RGSs and at the temperature where O VII is the dominant ion of oxygen, the low-T boundary is mostly set by the non-detection of K-shell transitions of O IV-O VI and M-shell transitions of Fe VII–Fe xVI, whereas the high-T boundary is set by the absence of the K-shell transitions of O vIII and Ne IX-Ne x and the L-shell transitions of Fe xvii-Fe xxiv.

We note that our temperature estimates depend on our assumption of gas in collisional ionization equilibrium, perturbed by the meta-galactic ionizing radiation field. This assumption may not be valid in particularly low-density regions, where the post-shock electron–ion relaxation timescale is longer than the Hubble time¹⁹, and could lead to an overestimation of the actual electron temperature¹⁹. However, even for a difference by a factor of 2 in electron temperature, the fraction of our tracer ion, O vII, at its peak temperature interval, $T\approx 5\times 10^5-2\times 10^6$ K, is not considerably affected by this effect (see, for example, Fig. 4 of ref. ¹⁹). This makes our oxygen column density estimates virtually independent of the exact ionization state of the gas.

The EWs of the detected O VII lines, together with the limits on the EWs of higher- and lower-ionization oxygen lines, constrain the total oxygen column densities of the X-ray absorbers and, modulo their metallicity (which we assume to be Z_{\odot} in the fit), also their equivalent H column densities.

Finally, we use our hybrid-ionization models¹⁷ to derive ionization fractions of H I, O VI and O VII in a given temperature interval and for a gas volume density of $n_e = 10^{-5}$ cm⁻³ (at densities $n_e > 10^{-4}$ cm⁻³, photoionization plays a negligible role, and the gas is virtually in collisional ionization equilibrium, whereas for $n_e \le 10^{-4}$ cm⁻³ the uncertainties in the ionization fractions of H I, O VI and O VII

are dominated by the uncertainties on the temperature of the absorber rather than its volume density).

HST-COS spectrum of 1ES 1553+113. 1ES 1553+113 was first observed with the HST-COS on 22 September 2009 for 3.1 ks and 3.8 ks, using the G130M (λ =1,135-1,480 Å) and G160M (λ =1,400-1,795 Å) gratings, respectively. This spectrum has a signal-to-noise ratio per $\Delta\lambda\approx0.08$ Å resolution element of SNRE \approx 23. This observation was published in ref. ²⁰, which reported the detection of 42 intervening IGM systems at $z\lesssim0.4$. A second spectrum was collected almost two years later, on 24 July 2011, with exposures of 6.4 ks (G130) and 8 ks (G160), which almost doubled the SNRE of the 2009 spectrum over its entire G130M+G160M bandpass.

Here we present a targeted analysis of the full HST-COS spectrum of 1ES 1553+113, aimed at searching specifically for H I or O VI counterparts to the two intervening O vII absorbers identified in the XMM-Newton RGS spectra. We do this by using the fitting package Sherpa of the Chandra Interactive Analysis of Observation (CIAO) software to model the normalized HST-COS spectrum with negative Gaussian functions. When lines are seen in the HST-COS spectrum, we leave all the Gaussian parameters (namely, position, width and normalization factor, or EW) free to vary in the fit and let the fitting routine find the best-fitting parameters by minimizing the statistics with the method *moncar*, which is included in Sherpa. When more than one transition from the same ion is present and the association is secure, we model the absorption lines by linking their relative positions and widths to the relative rest-frame positions. When lines are not seen, but an EW upper limit is needed for a line with position and width bounded by physically motivated limits (that is, the redshifts of the two intervening X-ray systems for the line positions, or their temperatures for the line widths), we set these limits as the minimum and maximum values of the Gaussian position and width during the fit, and let the fitting routine find the best-fitting Gaussian normalization within these limits. We then compute the 3σ upper limit on the normalization of such best-fitting Gaussian by using the Sherpa task conf.

When resolved lines of H I are available, we use the empirical correction of ref. 21 ($b_{\rm HI}^{\rm th} = \sqrt{2kT/m_p}$ and $b_{\rm HI}^{\rm th} \approx b_{\rm HI}^{\rm obs}/1.2$, where $b_{\rm HI}^{\rm th}$ and $b_{\rm HI}^{\rm obs}$ are the thermal and observed Doppler parameter, respectively, k is the Boltzmann constant and m_p is the proton mass) to derive the temperature of the FUV gas. Analogously, we use the inverse relationship to set boundary conditions to the width of a H I (or O VI, by factoring in the O/H mass factor) line, when evaluating the 3σ upper limit on the EW of such a line.

Following ref. 22 , we derive the ion column densities of FUV lines by running curve-of-growth analyses of the measured line EWs. When more than two transitions from the same ion are present, we check for common solutions in the ion column density versus the Doppler parameter plane (for example, Ly α and Ly β lines near System 2; see 'H I and O VI counterparts of X-ray systems 1 and 2') and then use the estimated temperature to further constrain the ion column density. H I and O vI counterparts of X-ray systems 1 and 2. The H I Lya counterpart to the X-ray System 1 falls just on the long-wavelength side of the strong Ni II $(\lambda = 1,741.55 \text{ Å})$ line from our Galaxy's interstellar medium. This Ni II line was misidentified in ref. ²⁰ as a H I Ly α absorber at z = 0.43261. In our re-analysis of the HST-COS spectrum of 1ES 1553+113, we simultaneously fit the six available Ni II lines ($\lambda = 1,317.22$ Å, $\lambda = 1,370.13$ Å, $\lambda = 1,454.84$ Å, $\lambda = 1,709.40$ Å, $\lambda = 1,741.55$ Å and $\lambda = 1,751.9$ Å). We then add a negative Gaussian to the model to search for a H I Ly α absorber at redshifts within $\pm 1\sigma$ from z_1^X and thermal width bounded by the $\pm 1\sigma$ uncertainty on T_1^X . The fitting routine finds a weak (2.3 σ) and broad ($b_{\rm HI}^{\rm obs} \approx 220~{\rm kms}^{-1}$) line (Extended Data Table 1), which we use to set an upper limit on the EW of the H I Ly α associated with System 1. Analogously, we use the boundary position and width conditions imposed by the redshift and temperature of System 1 to set an upper limit to the EW of the O vI ($\lambda = 1,031.93 \text{ Å}$) line of System 1 (Extended Data Table 1).

In System 2, two H I Ly α (Fig. 2b) and Ly β (Fig. 2c) absorbers are present at redshifts consistent with that of the O VII He α absorber (Fig. 2a), $z_{\rm FUV}=0.35383\pm0.00001$ and $z_{\rm FUV}=0.35642\pm0.00001$. However, the H I lines at $z_{\rm FUV}=0.35642$ are too narrow ($b_{\rm HI}^{\rm H}\approx32~{\rm kms}^{-1}$, implying $T\approx6\times10^4~{\rm K})^{21}$ to be even tentatively associated to the O VII-bearing gas. For the broader H I absorber at $z_{\rm FUV}=0.35383\pm0.00001$, we estimate $^{21}T_2^{\rm BHI}=(3.5^{+0.8}_{-0.7})\times10^5{\rm K}$, only marginally (2σ) consistent with the X-ray estimate. By combining the Doppler widths with the observed line EWs, we obtain $N_{\rm HII,2}^{\rm BHI}=(3.7\pm0.1)\times10^{14}{\rm cm}^{-2}$, $N_{\rm OVII,2}^{\rm BHI}=(5.2^{+3.5}_{-2.0})\times10^{15}{\rm cm}^{-2}$ and a 3σ upper limit $N_{\rm OVI,2}^{\rm BHI}<3.5\times10^{13}{\rm cm}^{-2}$. By factoring in the ionization fractions, we obtain a solid total H column density for the broader H I absorber: $N_{\rm H,2}^{\rm BHI}=(3.2\pm1.0)\times10^{20}{\rm cm}^{-2}$. However, the O columns, derived from O VII and O VI, are inconsistent ($>3\sigma$) with each other: $N_{\rm O,2}^{\rm BHI}$ (O VII) = ($7.7^{+5.7}_{-3.8}/\times10^{15}{\rm cm}^{-2}$ and $N_{\rm O,2}^{\rm FUV}$ (O VI) $<0.7\times10^{15}{\rm cm}^{-2}$. This, together with the X-ray-FUV temperature inconsistency, imply that (in the collisional ionization equilibrium hypothesis that we adopt here) neither of the two strong H I absorbers at redshifts consistent with $z_{\rm X}^{\rm X}$ can be physically associated to System 2. Following the same procedure as for System 1, we therefore set 3σ

upper limits to the EWs of the H I Ly α and O vI (λ = 1,031.93 Å) transitions (Extended Data Table 1).

FUV–X-ray diagnostics: refined temperature and metallicity. We use the upper limits on the EWs of the O vI (λ =1,031.93 Å) transition in System 1 and System 2 to refine the allowed O vI ionization fraction intervals, and so the lower boundaries of the allowed temperature intervals. Namely, we estimate an upper limit on the O vI ionization fraction by dividing the FUV-derived O vI column density upper limit by the 1σ lower boundary on the X-ray-derived oxygen column density: $f_{\rm OVI} = N_{\rm OVI}/N_{\rm O} < N_{\rm OVI}^{\rm FUV}/[N_{\rm O}^{\rm X} - (\Delta N_{\rm O}^{\rm X})]$. For System 1 this gives the revised temperature $T_{\rm I}^{\rm X} = (0.8-1.6) \times 10^6 {\rm K}$ and ionization fractions $f_{\rm I}^{\rm HI} = (2.3\pm1.0)\times10^{-7}, f_{\rm I}^{\rm OVI} = (0.0041\pm0.0018) {\rm and} f_{\rm I}^{\rm OVII} = (0.811\pm0.098).$ For System 2 we obtain a revised temperature $T_{\rm I}^{\rm X} = (0.5-1.4)\times10^6 {\rm K}$ and ionization fractions $f_{\rm I}^{\rm HI} = (4.0\pm2.5)\times10^{-7}, f_{\rm I}^{\rm OVI} = (0.019\pm0.016)$ and $f_{\rm I}^{\rm OVII} = (0.855\pm0.060).$

We then use these refined ionization corrections and the upper limits on the EWs of the H I Ly α transition to derive 3σ upper limits on the FUV-equivalent H column densities, and so the 3σ lower limits on the metallicity of the systems. We do this by first dividing the 3σ upper limit on the H I column density by the central value of the revised H I ionization fraction, $N_{\rm H}^{\rm FUV} = N_{\rm HI}^{\rm FUV}/f_{\rm HI}$, and then comparing this with the central value of the X-ray-estimated equivalent H column, $Z/Z_{\odot} = N_{\rm H}^{\rm X}/N_{\rm H}^{\rm FUV}$. As an upper limit on the mean metallicity of the systems, we instead assume the value $Z_{\rm ICM} = 0.2Z_{\odot}$ found in the peripheries (at r_{500}) of the intra-cluster medium 15 . However, it is possible that the metal distribution in the IGM is inhomogeneous, and in particular it has been argued that oxygen absorbers can arise from relatively over-enriched ($Z \approx 0.5Z_{\odot}$) regions 12,13 . Should this be the case for one or both of our O VII absorbers, this would affect (by linearly lowering it) our lower limit on the cosmological mass density of baryons in the WHIM (see 'Cosmological mass density').

Cosmological mass density. Following ref. ²³, we estimate the cosmological mass density of baryons in the $10^{5.7}$ K \leq $T \leq 10^{6.2}$ K WHIM by using the formula

$$\Omega_{\rm b}h^2(10^{5.7} \text{ K} \le T \le 10^{6.2} \text{ K}) = \left(\frac{1}{\rho_{\rm c}}\right) \left(\frac{m_p \sum_i N_{\rm H}^i}{(1-Y)d}\right)$$

where (1-Y) is the hydrogen mass fraction (whose inverse is taken to be \sim 1.3), $\rho_{\rm c}$ is the universe critical density, $N_{\rm H}^i$ (for i=1,2) is the estimated equivalent H column density for systems 1 and 2 and d is the available path length, which (after factoring in the reduction of available RGS bandpass due to both Galactic absorption lines and detector cool pixels) is $\Delta z=0.42$ or d=1,528 Mpc comoving.

Optical data of 1ES 1553+113. We observed the field of 1ES 1553+113 with the OSIRIS camera at the 10-m Gran Telescope Canarias (GTC). We performed a 4 × 4 mosaic observation centred on 1ES 1553+113 to cover a 30' × 30' field, in the 5 Sloan Digital Sky Survey (SDSS) bands, u', g', r', i' and z'. Our survey is flux-limited to $r' \approx 23.5$ (a factor of four deeper than the available SDSS data), corresponding to an absolute r' magnitude of -18.9 at z=0.5. We reduced the data using IRAF, along with the gtcmos package (http://www.inaoep.mx/~ydm/gtcmos/gtcmos.html). We detected galaxies and performed photometry using SExtractor. Following ref. 24 , we then derived photometric redshifts using photoraptor with a combination of the four band colours plus the pivot magnitude in the r' band. Our photometric redshifts have an accuracy of $\Delta z \approx 0.07$ in the interval $z \approx 0.15-0.6$.

The histogram in Extended Data Fig. 5 is built by considering photometrically identified galaxies in cylindrical volumes with a base radius of 0.5 Mpc and a line-of-sight depth of $\Delta z = 0.07$ at each redshift bin. Galaxies circled in yellow and cyan in Extended Data Figs. 6 and 7, respectively, are photometrically identified galaxies in cylindrical volumes with base radii of 0.5 Mpc and 1.75 Mpc and a line-of-sight depth of $\Delta z = 0.07$, centred at the redshifts of System 1 and System 2. According to our photometric redshifts, both System 1 and System 2 sit at the centre of a large concentration of galaxies: (12, 54) and (8, 72) galaxies with r'>23.5 are found in cylindrical volumes with base radii of (0.5, 1.5) Mpc and a line-of-sight depth of $\Delta z = 0.07$ in the redshift bins of System 1 (where only \sim (3, 27) are expected 11) and System 2 (only \sim (2, 18) expected 11), respectively.

We also have a number of spectroscopic redshifts taken with OSIRIS-MOS at GTC and with the GMOS at the Gemini North Telescope. Data have been reduced and analysed with IRAF. Redshifts are present for 44 galaxies in the 5.5′ × 5.5′ field around 1ES 1553+113 (35 new measurements and 6 from ref. 25). For System 1, 13 of our photometrically identified galaxies have spectroscopic redshift and for 8 of them the identification is confirmed (that is, redshifts within $\pm 900~\rm km~s^{-1}$ from the absorber; solid thick circles in Extended Data Fig. 6). For System 2 only 4 of our photometrically identified galaxies have spectroscopic redshifts and only one of these four identifications is confirmed (solid thick circle in Extended Data Fig. 7), at a velocity of $+370~\rm km~s^{-1}$ from the absorber's redshift.

On the implausibility of intrinsic absorption. Given the proximity of our two systems with the upper limit $z \lesssim 0.48$ that we estimate for the redshift of our target,

we cannot rule out that these absorbers are imprinted by material intrinsic to the blazar environment and outflowing from this environment at speeds not exceeding $\sim\!0.05c\!-\!0.12c$. However, a number of reasons (also discussed for an analogous case in ref. 17 , section 7.6.1), make this scenario implausible. Here we review some of these reasons for the specific cases of System 1 and System 2.

Although photoionized outflows are commonly seen in type-1 Seyferts and quasars (the so called 'warm absorbers'; for example, ref. 26 and references therein), these systems are always seen in multiple species (both in the FUV and X-rays) owing to the smoother ion fraction distribution in photoionized versus collisionally ionized gas (compare, for example, the top and bottom panels in Figs. 1 and 2 of ref. 27). By contrast, our System 1 and System 2are only seen in O VII, suggesting that electron—ion collisions are the main mechanism of ionization of this gas. Moreover, warm absorbers have typical outflow velocities of only few hundreds to few thousands of kilometres per second, whereas even for our System 1 the implied outflow velocity could be as high as 15,000 km s $^{-1}$. Finally, warm absorbers have equivalent hydrogen column densities in the range $N_{\rm H}\approx 10^{20}$ – 10^{23} cm $^{-2}$, at least an order of magnitude higher than those observed here 26 .

Indeed, no such system has ever been confirmed in blazars, and it is exactly the intrinsic featurelessness of their spectra (together with their relatively high X-ray fluxes) that makes blazars particularly suited for IGM-absorption X-ray experiments (for example, ref. 2 and references therein). Early reports of X-ray absorbers in blazars 28,29 have not been confirmed by later, higher-spectral-resolution observations $^{30-32}$. The only absorption lines detected in the high-resolution, high-signal-to-noise-ratio X-ray spectra of about a dozen of blazars are either from our own Galaxy's disk or halo 22,33,34 or claims of intergalactic WHIM (see ref. 2 and references therein). The only two exceptions reported are a misidentification of an O VII He α line at a redshift consistent with that of Mkn $421^{17,35}$, which is instead a K β transition of O II from our own Galaxy 22 , and a transient O VIII K α absorber at the redshift of the blazar H $2356-309^{36}$.

The O VIII K α absorber reported in ref. ³⁶ has an O VIII column two orders of magnitude larger than those reported here for our two O VII absorbers and, perhaps more importantly, it has a transient nature, appearing in only one out of the five consecutive ~80 ks Chandra observations of H 2356-309 performed in September 2008. This is not surprising; any gaseous material intrinsic to the AGN environment must experience strong photoionization by the quasar's radiation field, which in X-rays varies on timescales as short as few hundred seconds. Consistently, the ionization degree of warm absorbers is often seen to vary on timescales from few kiloseconds³⁷ to months³⁸. By contrast, no change is seen in our absorbers over the two years between the first and the second half of our observing campaign (see Extended Data Fig. 4 and Extended Data Table 2, where we also report the EWs of the strong Galactic line of N I during the two epochs for comparison), despite the orders-of-magnitude change commonly experienced by the beamed X-ray luminosity of our target³⁹. The lack of variability of our O VII absorbers would require electron volume densities of $n_e \lesssim 10^2 \, \mathrm{cm}^{-3}$ for these systems—at least two orders of magnitude lower than the lowest limit estimated for warm absorbers³⁸. Therefore, an intrinsic AGN outflow origin for our System 1 (and System 2) seems unlikely.

Such low densities are consistent with the disk or moderately extended halo of the blazar's host galaxy. However, in such a case, only our System 1 could be associated to the blazar's host galaxy (whose redshift would then coincide with that of System 1) and strong O 1 and O 11 K α absorbers should also be seen (as in our own Galaxy; for example, ref. 22 and Extended Data Fig. 1), but are not.

We conclude that a much more plausible explanation for our System 1 and System 2 is intervening absorption by diffuse WHIM or the CGM of intervening galaxies.

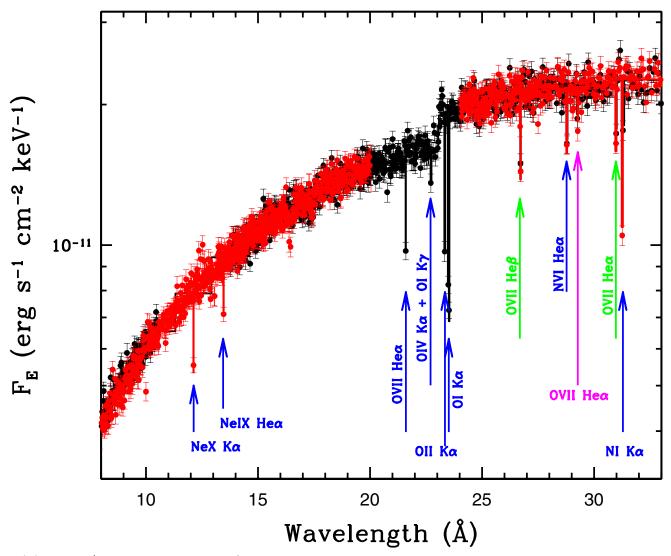
Ruling out absorption by a thick disk of an intervening galaxy. Our O VII observations for systems 1 and 2 could in principle be due to absorption by highly ionized material in the thick disk or halo of an intervening galaxy with an impact parameter <100 kpc (for the brightest spirals). In this case, the derived equivalent H column densities should not be used as described in 'Cosmological mass density' to derive the cosmological mass density of hot baryons in the Universe. However, there is no r' > 23.5 galaxy with confirmed impact parameter < 100 kpc from our two absorbers. Moreover, even if small galaxies, fainter than our survey's limit, were present at such small impact parameters, their thick disks or haloes would also contain a large amount of cool matter with high fractions of neutral and mildly ionized metals. Strong O I and O II absorption is seen, for example, at high galactic latitude in our own Galaxy²², as well as around the disk (galaxy-absorber impact parameters of \sim 6–100 kpc) of external galaxies⁴⁰. In ref. ²², comparison between lines of sight through the disk of our Galaxy and high-galactic-latitude lines of sight shows that a cool ($T \approx 3,000 \text{ K}$) ionized medium traced by O I and O II absorbers (also seen here in the spectrum of 1ES 1553+113, together with N I; Fig. 1 and Extended Data Fig. 1) fills both the disk and an extended thick disk or halo of our Galaxy.

Thus, if our System 1 and System 2 were analogous to our local O VII absorbers, and this highly ionized gas were confined in the thick disk or halo of small intervening galaxies, this gas would also probably co-exist with much cooler gas and would create both O VII and strong O I and O II absorption lines in the spectrum of 1ES 1553+113, which are not seen.

Data availability. The entire RGS and COS data used in this work are available in the public XMM-Newton and HST archives, namely, the XMM-Newton Science Archive (http://nxsa.esac.esa.int/nxsa-web/#home) and the Mikulsky Archive for Space Telescopes (MAST; https://archive.stsci.edu/hst/). In particular, we used the XMM-Newton datasets: 0094380801, 0656990101, 0727780101, 0727780201, 0727780301, 0761100101, 0761100201, 0761100301, 0761100401, 0761100701, 0761110101, 0790380501, 0790380601, 0790380801, 0790380901, 0790381001, 0790381401 and 0790381501. We also used the HST-COS G130 and G160 datasets LB4R02010, LB4R02020, LB4R02030, LB4R02040, LB4R02050, LB4R02060, LB4R02070, LB4R02080, LBG803070, LBG803080, LBG803090, LBG803090, LBG8030F0, LBG8030G0 and LBG8030H0, all publicly available at MAST.

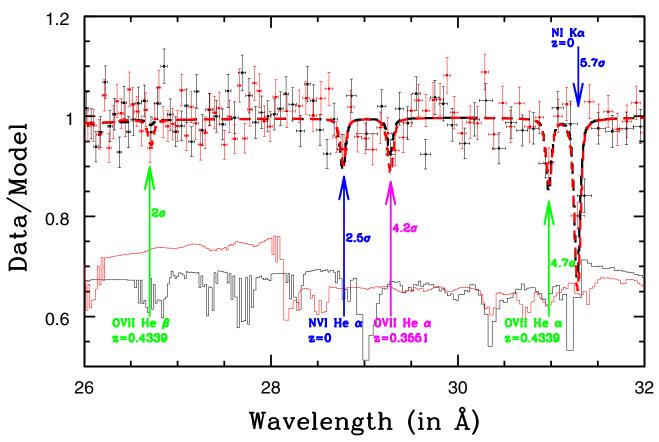
Code availability. The models, hybrid-ionization spectral code and Monte Carlo algorithms used in this work are available upon request from the corresponding author.

- H.E.S.S. Collaboration. The 2012 flare of PG 1553+113 seen with H.E.S.S. and Fermi-LAT. Astrophys. J. 802, 65 (2015).
- Nicastro, F. et al. Chandra detection of the first X-ray forest along the line of sight to Markarian 421. Astrophys. J. 629, 700–718 (2005).
- Asplund, M., Grevesse, N., Sauval, A. J. & Scott, P. The chemical composition of the Sun. Annu. Rev. Astron. Astrophys. 47, 481–522 (2009).
- Yoshida, N., Furlanetto, S. R. & Hernquist, L. The temperature structure of the warm-hot intergalactic medium. Astrophys. J. 618, L91–L94 (2005).
- Danforth, C. W., Keeney, B. A., Stocke, J. T., Shull, J. M. & Yao, Y. Hubble/COS observations of the Lyα forest toward the BL Lac object 1ES 1553+113.
 Astrophys. J. 720, 976–986 (2010).
- Danforth, C. W., Stocke, J. T. & Shull, J. M. Broad H I absorbers as metallicityindependent tracers of the warm–hot intergalactic medium. Astrophys. J. 710, 613–633 (2010).
- Nicastro, F. et al. X-ray detection of warm ionized matter in the Galactic halo. Mon. Not. R. Astron. Soc. 457, 676–694 (2016).
- 23. Schaye, J. Model-independent insights into the nature of the Ly $_{\alpha}$ forest and the distribution of matter in the Universe. Astrophys. J. **559**, 507–515 (2001).
- Brescia, M., Cavuoti, S., Longo, G. & De Stefano, V. A catalogue of photometric redshifts for the SDSS-DR9 galaxies. Astron. Astrophys. 568, A126 (2014).
- Prochaska, J. X., Weiner, B., Chen, H.-W., Cooksey, K. L. & Mulchaey, J. S. Probing the IGM/galaxy connection. IV. The LCO/WFCCD galaxy survey of 20 fields surrounding UV-bright quasars. Astrophys. J. Suppl. Ser. 193, 28 (2011).
- Reeves, J. N. et al. A high resolution view of the warm absorber in the quasar MR 2251-178. Astrophys. J. 776, 99 (2013).
- Nicastro, F., Fiore, F., Perola, G. C. & Elvis, M. Ionized absorbers in active galactic nuclei: the role of collisional ionization and time-evolving photoionization. *Astrophys. J.* 512, 184–196 (1999).
- Kruper, J. & Canizares, C. R. A sharp X-ray absorption feature in the BL Lac Object PKS 2155-304. Bull. Am. Astron. Soc., 14 933 (1982).
- Madejski, G. M., Mushotzky, R. F., Weaver, K. A., Arnaud, K. A. & Urry, C. M. A ubiquitous absorption feature in the X-ray spectra of BL Lacertae objects. Astrophys. J. 370, 198–204 (1991).
- Nicastro, F. et al. Chandra discovery of a tree in the X-ray forest toward PKS 2155-304: the local filament? Astrophys. J. 573, 157–167 (2002).
- Fang, T., Sembach, K. R. & Canizares, C. R. Chandra detection of local O VII Heα absorption along the sight line toward 3C 273. Astrophys. J. 586, L49–L52 (2003)
- Cagnoni, I., Nicastro, F., Maraschi, L., Treves, A. & Tavecchio, F. A view of PKS 2155-304 with XMM-Newton reflection grating spectrometers. *Astrophys. J.* 603, 449–455 (2004).
- Bregman, J. N. & Lloyd-Davies, E. J. X-ray absorption from the Milky Way halo and the local group. Astrophys. J. 669, 990–1002 (2007).
- Nicastro, F., Senatore, F., Krongold, Y., Mathur, S. & Elvis, M. A distant echo of Milky Way central activity closes the Galaxy's baryon census. Astrophys. J. 828, L12 (2016).
- 35. Rasmussen, A. P. et al. On the putative detection of z > 0 X-ray absorption features in the spectrum of Mrk 421. Astrophys. J. **656**, 129–138 (2007).
- Fang, T., Buote, D. A., Humphrey, P. J. & Canizares, C. R. Detection of a transient X-ray absorption line intrinsic to the BL Lacertae object H 2356-309. Astrophys. J. 731, 46 (2011).
- Krongold, Y. et al. The compact, conical, accretion-disk warm absorber of the Seyfert 1 galaxy NGC 4051 and its implications for IGM-galaxy feedback processes. Astrophys. J. 659, 1022–1039 (2007).
- Krongold, Y. et al. Suzaku monitoring of the Seyfert 1 galaxy NGC 5548: warm absorber location and its implication for cosmic feedback. Astrophys. J. 710, 360–371 (2010).
- Ackermann, M. et al. Multiwavelength evidence for quasi-periodic modulation in the gamma-ray blazar PG 1553+113. Astrophys. J. 813, L41 (2015).
- Lehner, N. et al. The bimodality distribution of the cool circumgalactic medium at z ≤ 1. Astrophys. J. 770, 138 (2013).



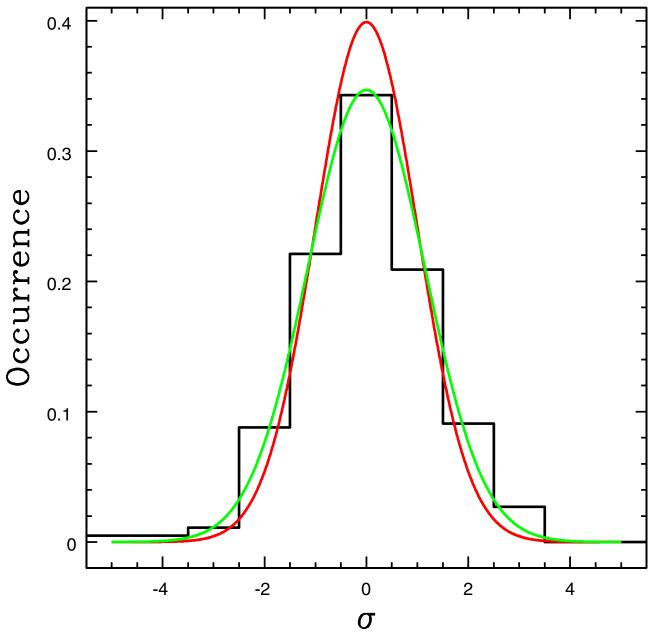
Extended Data Fig. 1 | XMM-Newton RGS spectra of 1ES 1553+113. Broadband, unfolded RGS1 (black points; 1σ error bars) and RGS2 (red points; 1σ error bars) spectra (in bins with a signal-to-noise ratio per bin \geq 20) and best-fitting models (black and red histograms) for the blazar

1ES 1553+113, in physical units. Blue arrows mark Galactic absorption lines; green and magenta arrows indicate absorption lines from our WHIM System 1 and System 2, respectively. F_E , source flux (power per unit area and energy).



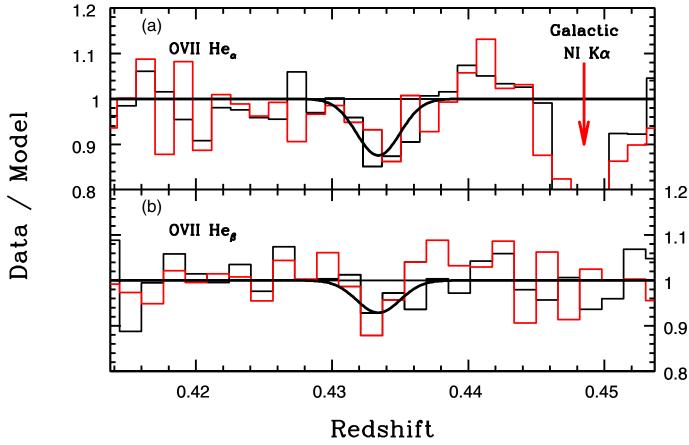
Extended Data Fig. 2 | Normalized XMM-Newton RGS spectra of 1ES 1553+113. Normalized raw RGS1 (black points; 1σ error bars) and RGS2 (red points; 1σ error bars) data (in bins with a signal-to-noise ratio per bin ≥ 30) of the blazar 1ES 1553+113, in the wavelength interval $\lambda = 26-32$ Å. Thick dashed curves are the RGS1 (black) and RGS2 (red) best-fitting models, folded through the response functions of the RGSs.

Thin solid curves at the bottom of the graph are RGS1 (black) and RGS2 (red) effective areas (in arbitrary units), showing instrumental features due to cool pixels in the dispersing detectors. Of the five absorption lines shown, only the weak O vII He β at $z^X = 0.4339$, and only in RGS1, can be affected by the presence of an instrumental feature.



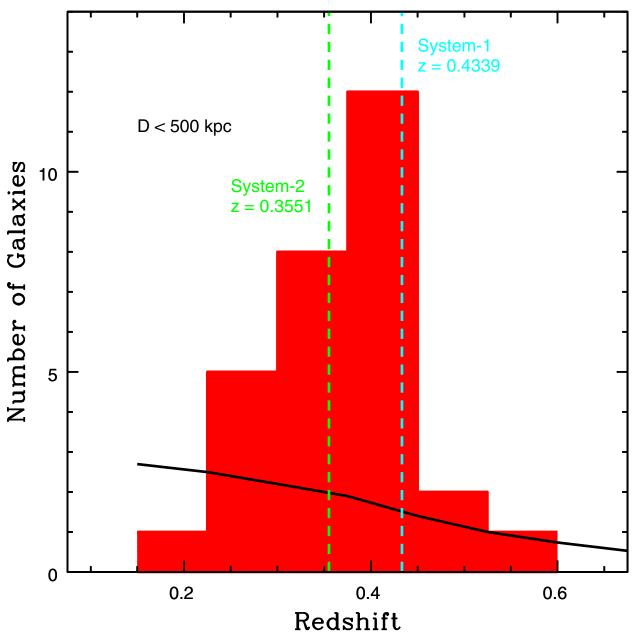
Extended Data Fig. 3 | Assessing systematic errors in the RGS spectrum of 1ES 1553+113. Outcome of a Monte Carlo procedure consisting of 10,000 evaluations of the single-line statistical significance of an unresolved (that is, broadened to the instrument line spread function) Gaussian added to the best-fitting continuum-plus-line model of the RGS spectrum of 1ES 1553+113 (see Methods). At each run, the line position

is frozen at a random position in the range 8–33 Å, and its EW is allowed to vary freely from negative (emission) to positive (absorption). The data distribution is symmetric but slightly flatter than the red curve that shows the expected normal distribution for a standard deviation of unity. The normal distribution that best fits our Monte Carlo results has a standard deviation of 1.15 (green curve).



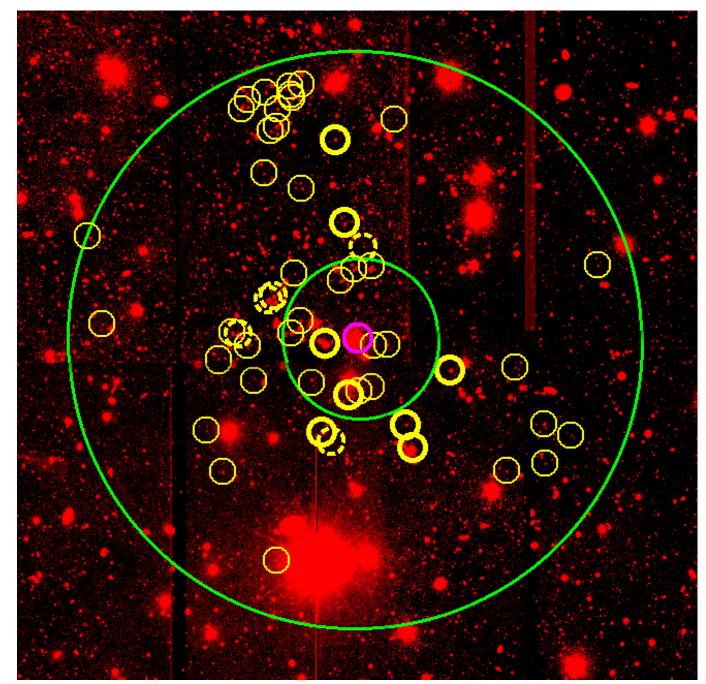
Extended Data Fig. 4 | Constancy of the z=0.4339 O VII absorber. a, b, RGS spectra of 1ES 1553+113 obtained in the 2015 (black) and 2017 (red) observations for 800 ks and 950 ks, respectively, centred around

the O VII He α (a) and He β (b) transitions of System 1. The two lines are consistent with no variability between the two epochs, within their 1σ errors (Extended Data Table 2).



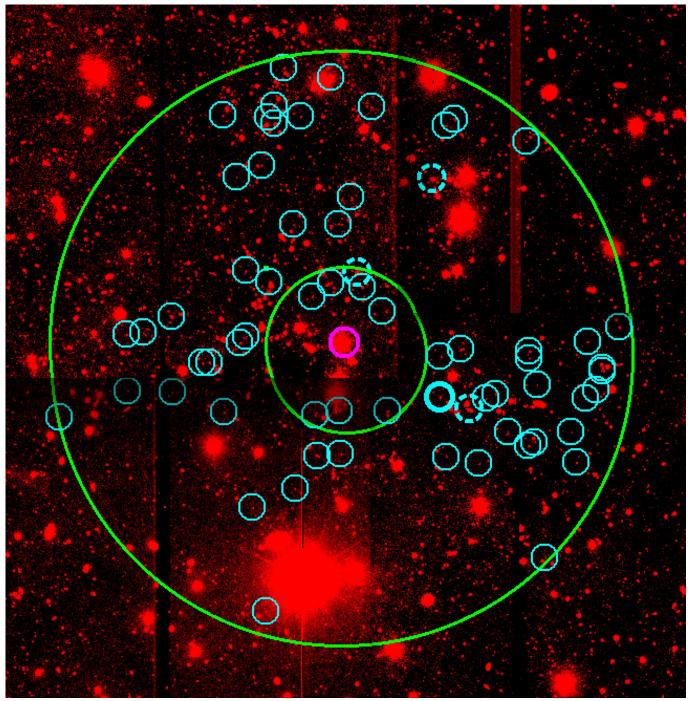
Extended Data Fig. 5 | **Galaxy photometric redshifts.** Histogram of photometric redshifts of r'>23.5 galaxies within cylindrical volumes with base radii of 500 kpc at each redshift interval and a line-of-sight depth of $\Delta z=0.075$ (at 1σ redshift accuracy), centred on the line of sight to the blazar 1ES 1553+113. The black curve is the average number of galaxies with r'>23.5 expected within the explored volumes at each

redshift bin, based on the galaxy $\log N - \log S$ from ref. ¹¹ (where N is the number of galaxies with surface brightness $\geq S$ and we use a common but conservative B-r'=1 for all galaxy types, where B is the galaxy's B magnitude). Vertical dashed green and blue lines indicate the redshifts of System 1 and System 2, respectively.



Extended Data Fig. 6 | r'>23.5 galaxies surrounding WHIM System 1. OSIRIS-GTC mosaic (in band r') of the $12'\times12'$ field surrounding the line of sight to 1ES 1553+113 (indicated by the magenta circle). Green circles have radii of 500 kpc and 1.75 Mpc at z=0.4125. Thin yellow circles

highlight the positions of galaxies with photometric redshift estimates within the $z\!=\!0.375\!-\!0.450$ interval, whereas thicker circles highlight spectroscopic redshifts (solid, confirmed; dashed, unconfirmed).



Extended Data Fig. 7 | r' > 23.5 galaxies surrounding WHIM System 2. Same as Extended Data Fig. 5, but for System 2, in the z = 0.300-0.375 interval.

Extended Data Table 1 \mid Absorption lines of System 1 and System 2

Wavelength	Id	Redshift b		EW _{obs}	Significance
(Å)			(km s ⁻¹)	(mÅ)	(σ)
30.975±0.017	OVII He-α	0.4339±0.0008	NA	14.7±3.1	4.1–4.7
26.69±0.09	OVII He-β	0.4326±0.0048	NA	$4.4^{+2.7}_{-2.2}$	1.7-2.0
1742.18-1744.24	HΙ Ly-α	0.4331-0.4347	90-220	<110	3.0u.l.
1478.86-1480.51	$\begin{array}{c} OVI \\ 2s2p_{1/2} \end{array}$	0.4331-0.4347	55-128	<30	3.0 u.l.
$29.27^{+0.01}_{-0.03}$	OVII He-α	$0.3551^{+0.0003}_{-0.0015}$	NA	$10.5^{+2.9}_{-2.5}$	3.7-4.2
1645.53-1647.72	HΙ Ly-α	0.3536-0.3554	96-190	<98	3.0 u.l.
1396.82-1398.68	$\begin{array}{c} OVI \\ 2s2p_{1/2} \end{array}$	0.3536-0.3554	57-113	<49	3.0 u.l.

Id, identification; u.l., upper limit; EW_{obs}, observed EW.

Extended Data Table 2 \mid Lack of variability of the O VII absorbers over two years

Epoch	Transition	System	EW _{obs} (in mÅ)
1	OVII He-α	1	17±5
2	OVII He-α	1	15±5
1	OVII He-β	1	7±4
2	OVII He-β	1	8±3
1	OVII He-α	2	8±5
2	OVII He-α	2	12±4
1	ΝΙ Κα	Gal	39±7
2	ΝΙ Κα	Gal	37±6

Gal, Galactic.



Extended Data Table 3 \mid Physics and chemistry of System 1 and System 2

System	T (10 ⁶ K)	N ₀ (10 ¹⁵ cm ⁻²)	$N_{\rm H}({\rm Z/Z}_{\odot})^{-1}$ (10 ¹⁹ cm ⁻²)	Z (Z _☉)
1	1.2±0.4	$7.8^{+3.9}_{-2.4}$	$1.6^{+0.8}_{-0.5}$	≥0.1
2	0.95±0.45	$4.4^{+2.4}_{-2.0}$	$0.9^{+0.5}_{-0.4}$	≥0.1



Gate-tunable frequency combs in graphene-nitride microresonators

Baicheng Yao 1,2,7,11* , Shu-Wei Huang 1,8,11* , Yuan Liu 3,9,11 , Abhinav Kumar Vinod 1,11 , Chanyeol Choi 1 , Michael Hoff 1 , Yongnan Li 1 , Mingbin Yu 4,10 , Ziying Feng 5 , Dim-Lee Kwong 4,6 , Yu Huang 3 , Yunjiang Rao 2 , Xiangfeng Duan 5* & Chee Wei Wong 1*

Optical frequency combs, which emit pulses of light at discrete, equally spaced frequencies, are cornerstones of modern-day frequency metrology, precision spectroscopy, astronomical observations, ultrafast optics and quantum information 1-7. Chipscale frequency combs, based on the Kerr and Raman nonlinearities in monolithic microresonators with ultrahigh quality factors⁸⁻¹⁰, have recently led to progress in optical clockwork and observations of temporal cavity solitons¹¹⁻¹⁴. But the chromatic dispersion within a laser cavity, which determines the comb formation 15,16, is usually difficult to tune with an electric field, whether in microcavities or fibre cavities. Such electrically dynamic control could bridge optical frequency combs and optoelectronics, enabling diverse comb outputs in one resonator with fast and convenient tunability. Arising from its exceptional Fermi-Dirac tunability and ultrafast carrier mobility¹⁷⁻¹⁹, graphene has a complex optical dispersion determined by its optical conductivity, which can be tuned through a gate voltage^{20,21}. This has brought about optoelectronic advances such as modulators^{22,23}, photodetectors²⁴ and controllable plasmonics^{25,26}. Here we demonstrate the gated intracavity tunability of graphenebased optical frequency combs, by coupling the gate-tunable optical conductivity to a silicon nitride photonic microresonator, thus modulating its second- and higher-order chromatic dispersions by altering the Fermi level. Preserving cavity quality factors up to 106 in the graphene-based comb, we implement a dual-layer iongel-gated transistor to tune the Fermi level of graphene across the range 0.45-0.65 electronvolts, under single-volt-level control. We use this to produce charge-tunable primary comb lines from 2.3 terahertz to 7.2 terahertz, coherent Kerr frequency combs, controllable Cherenkov radiation and controllable soliton states, all in a single microcavity. We further demonstrate voltage-tunable transitions from periodic soliton crystals to crystals with defects, mapped by our ultrafast second-harmonic optical autocorrelation. This heterogeneous graphene microcavity, which combines singleatomic-layer nanoscience and ultrafast optoelectronics, will help to improve our understanding of dynamical frequency combs and ultrafast optics.

Figure 1a–c shows the concept and fabrication of our graphene gate-tunable Kerr frequency comb with source–drain and top gating. This is further detailed in the Methods and in Supplementary Information sections 1 and 2. To ensure transparency and minimal effect on the resonator quality factor (Q) for coherent comb generation, we top-gate the interacting graphene to pull the Fermi level up to 0.6 eV for reduced photon absorption in the nearly massless Dirac cone. An ion-gel capacitor is implemented on top of the graphene monolayer²⁷. The electric double layer in the ionic liquid provides a capacitance up to about $7.2\,\mu\text{F cm}^{-2}$; this high value enables high doping control

and comb tunability with a few-volt-level gating. This is important to produce sharp modulation of the cavity chromatic dispersion while keeping the cavity loss low. In addition to the optimized 300-nm gap between the Si₃N₄ waveguide and the graphene layer, we optimize the planar interaction length of the arc in which the graphene overlaps the nitride resonator to be about $80\,\mu\text{m}.$ The grey ring shown in Fig. 1a is the nitride resonator. This offers substantial tunability of the frequency comb combined with minimal graphene absorption losses. Figure 1d plots the computed optical group velocity dispersion (β_2) and the computed third-order dispersion (β_3) for tuned Fermi levels from 0.2 eV to 0.8 eV of the graphene monolayer. For each Fermi level, we note the wavelength oscillations in both β_2 and β_3 , arising from the lifetime of the carrier relaxation oscillations in graphene captured in the resonance of the monolayer sheet conductivity. As a result, the graphene β_2 can be tuned from anomalous to normal dispersion and then back to anomalous by means of the gate voltage, which is important for nonlinear phase-matching tunability. This enables wide and tunable frequency comb generation in the graphene-based microresonator (GMR). Based on the modelled overall graphene β_2 and β_3 , we model the heterogeneous microresonator for Kerr frequency comb generation. Figure 1e shows the temporal map of the comb dynamics in the GMR, obtained by Lugiato–Lefever equation (LLE) modelling. At $E_{\rm F}$ = 0.2 eV, the Qfactor is low, and hence there is no comb generation. At $E_{\rm F}$ = 0.5 eV, the GMR has $Q \approx 8 \times 10^5$, $\beta_2 \approx -50 \, \text{fs}^2 \, \text{mm}^{-1}$ and $\beta_3 \approx 0$, resulting in slow comb generation. At $E_{\rm F} = 0.8$ eV, we observe rapid generation of a full comb in the numerical model, for $Q > 1 \times 10^6$, $\hat{\beta}_2 \approx -30 \,\text{fs}^2 \,\text{mm}^$ and $\beta_3 \approx -400 \, \text{fs}^3 \, \text{mm}^{-1}$.

Figure 2a shows the electrical tuning performance of graphene in the GMR. For a fixed source–drain voltage $V_{\rm SD}=10\,{\rm mV}$, the source–drain current $I_{\rm SD}$ is tuned with the gate voltage $V_{\rm G}$. When $V_{\rm G}$ reaches 2.4 V, $I_{\rm SD}$ has a minimum of 6.5 μ A. Here the carrier density of the graphene monolayer reaches the Dirac point. When $V_{\rm G}$ is less than 2.4 V, graphene is p-doped. In cyclic $V_{\rm G}$ tuning, a hysteresis loop is observed, owing to electronic trapping. The corresponding gate-tunable Fermi energy $|E_{\rm F}|=\hbar|v_{\rm F}|(\pi N)^{-1/2}$ (equation from ref. 28) is plotted in the bottom panel of Fig. 2a and noted to be proportional to $(V_{\rm G})^{1/2}$; here N is the carrier density, while v_F indicates the Fermi velocity. In our experiment, we tune $V_{\rm G}$ in the range $-2\,{\rm V}$ to 0 V, thereby controlling the graphene $|E_{\rm F}|$ between 0.65 eV and 0.45 eV. For $V_{\rm G}=0\,{\rm V}$, the graphene monolayer in our GMR is already heavily doped, which allows dispersion tuning with low loss.

Figure 2b maps the calculated real and imaginary parts of the GMR, varying with $|E_{\rm F}|$ and wavelength λ . In the two maps, the blue curves denote the boundary where dispersion abruptly changes, and the yellow curve denotes the low-loss region. In our measurement, we apply a high-power continuous-wave pump at 1,600 nm. At this wavelength,

¹Fang Lu Mesoscopic Optics and Quantum Electronics Laboratory, University of California, Los Angeles, CA, USA. ²Key Laboratory of Optical Fiber Sensing and Communications (Education Ministry of China), University of Electronic Science and Technology of China, Chengdu, China. ³Department of Materials Science and Engineering, University of California, Los Angeles, CA, USA. ⁴Institute of Microelectronics, Singapore, Sing

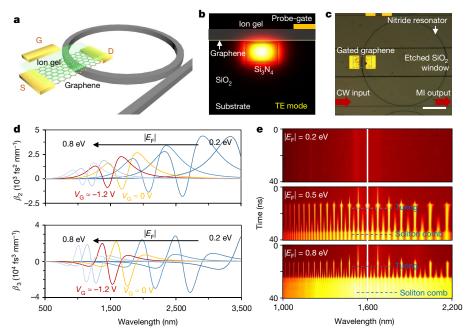


Fig. 1 | Conceptual design and implementation of the gate-tunable graphene–nitride heterogeneous microcavity. a, Schematic architecture of the GMR, with the silicon nitride indicated in grey. A graphene/ion-gel heterostructure is incorporated in the nitride microresonator. b, Electric-field distribution of the graphene–nitride heterogeneous waveguide, with a $\rm Si_3N_4$ cross-section of $\rm 1.2\times0.8\,\mu m^2$. The distance between the $\rm Si_3N_4$ waveguide and the graphene layer is $\rm 100$ nm. The graphene and the top-gate probe are separated by $\rm 1\,\mu m$ with the interlayer ion-gel capacitor. In this structure, transverse electric (TE) mode is applied. c, Optical micrographs show the bus waveguide (red arrows), ring resonator and Au/Ti metallized patterns. An etched window is designed

to ensure both graphene–light interaction and reduced propagation loss. Here the graphene-covered area is marked by the grey dashed box; the etched window label refers to the whole horizontal area between the two central lines. CW, continuous wave; MI: modulated intensity. Scale bar, $100\,\mu\text{m}$. **d**, Calculated group velocity dispersion and third-order dispersion of graphene, depending on its Fermi level. Here, the curves with $|E_F|=0.5\,\text{eV}$ and $|E_F|=0.6\,\text{eV}$, corresponding to the experimental conditions, are highlighted in yellow and red respectively. **e**, Simulated Kerr comb dynamics in the GMR, with different dispersion curves determined by the graphene Fermi level.

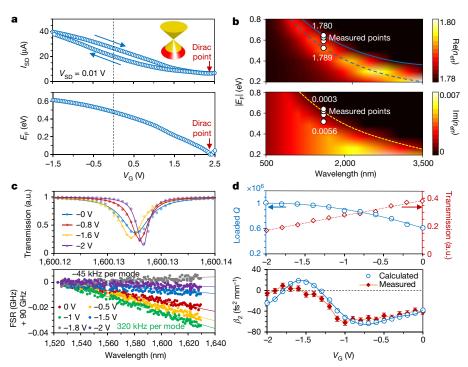


Fig. 2 | Gate-tuning the graphene microring resonator. a, Electronic measurement of the graphene/ion-gel capacitor. At a source–drain voltage $V_{\rm SD}=10$ mV, the correlation between $V_{\rm G}$ and $I_{\rm SD}$ shows the Dirac point position and tunable Fermi level of the graphene layer. b, Theoretically modelled $n_{\rm eff}$ of the GMR as a function of Fermi levels and optical wavelengths, in which the dispersion and Q can be deduced from the real and imaginary components. Measured data points are shown in white, at a wavelength of 1,600 nm, with $|E_{\rm F}|$ from 0.5 eV to 0.7 eV.

c, Measured transmissions (top panel) and mode FSR (bottom panel; dots, measured; curves, linear fitting) of the GMR, under gate voltages $V_{\rm G}$ from 0 V to -2 V. d, Tuned Q factor and dispersion, under various $V_{\rm G}$. The Q factor increases from 6×10^5 to 1×10^6 as the group velocity dispersion is controlled between $-62~{\rm fs^2~mm^{-1}}$ and $+9~{\rm fs^2~mm^{-1}}$. Error bar is the measurement uncertainty estimated from FSR measurements under the same condition.

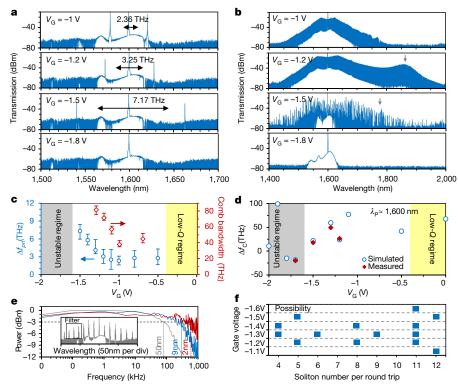


Fig. 3 | Observations of the gate-tunable graphene Kerr frequency combs. a, Primary comb lines at controlled gate voltages and Fermi levels of graphene. b, Full frequency combs generated under gate voltages of -1 V, -1.2 V, -1.5 V and -1.8 V. Here the launched pump power is fixed at 34.5 dBm. Kerr combs are generated by fine adjustment of the pump wavelength. Peaks of the Cherenkov radiation are marked by the grey arrows. c, Gate voltage tunes not only the primary comb line locations (blue circles) but also the full comb bandwidth (red diamonds).

d, Frequency spacing between the continuous-wave pump and the Cherenkov radiation, which is proportional to β_2/β_3 . **e**, 3-dB modulation bandwidths of 80 kHz, 200 kHz and 600 kHz are demonstrated by using optical filters with passband bandwidths of 50 nm, 9 nm and 2 nm respectively. The modulation speed is currently bounded by the ion-gel capacitance. **f**, Statistical distribution of the measured soliton states, with the same experimental parameters except for $V_{\rm G}$ which is tuned from $-1.1\,{\rm V}$ to $-1.6\,{\rm V}$.

when we tune $|E_{\rm F}|$ from 0.45 eV to 0.65 eV, the effective refractive index $n_{\rm eff}$ is controlled from 1.789 + 0.058i to 1.781 + 0.001i. Figure 2c shows the measured transmission and free spectral range (FSR; the wavelength spacing between successive maxima) dependences of the GMR, at different gate voltages. In this measurement, a broadband tunable laser serves as the light source at less than 10 mW, below the comb generation threshold. For a selected resonance around 1,600 nm, when $V_{\rm G}$ is tuned from 0 V to -2 V, the extinction ratio increases from 63% to 84%, and the resonance linewidth decreases from 3.1 pm to 1.6 pm. The mode deviation from equidistance, $D_{\rm FSR} = -\beta_2 c (2\pi f_{\rm FSR})^2/n_{\rm eff}$, is 320 kHz per mode for $V_{\rm G} \approx -1$ V (anomalous dispersion) but -45 kHz per mode for $V_{\rm G} \approx -1$ 8 V (normal dispersion), where c is the light velocity in vacuum and $f_{\rm FSR}$ is the frequency range of the FSR¹³. More details are shown in Extended Data Fig. 1.

Figure 2d shows the gate-tuning performance of the GMR. When the gate voltage is between 0 V and -2 V, the Fermi level remains higher than 0.4 eV, and thus graphene linear absorption in our working spectral range around 1,600 nm is strongly inhibited by Pauli blocking. As a result, the loaded Q factor of the GMR increases from about 6×10^5 to 10^6 , enabling comb generation under a 1-W pump, which is critical for both protecting the graphene monolayer from damage and stabilizing the frequency combs. We also note that Q-factor deterioration is induced by both the etching process and the linear absorption of the graphene heterostructure (Extended Data Fig. 2). For applications that require a higher Q factor, other 2D materials such as transition metal dichalcogenides with intrinsic bandgaps (for example WSe2) could be used to construct the heterogeneous microcavities²⁹. Simultaneously, the dispersion of the resonator is dynamically tuned, varying continuously from $-62 \,\mathrm{fs^2} \,\mathrm{mm^{-1}}$ anomalous dispersion to $+9 \,\mathrm{fs^2} \,\mathrm{mm^{-1}}$ normal dispersion. The group velocity dispersion tuning mainly results from

the graphene's Dirac–Fermi dynamics³⁰, with smaller contributions from the ion transport and thermal effects in the ion gel.

Next, we pump the GMR with 2-W continuous-wave laser power, with the primary comb lines (the strongest frequency combs generated from modulation instability initiation) shown in Fig. 3a under different gate voltages. For applied $V_G = -1 \text{ V}$, -1.2 V and -1.5 V, the frequency offsets between the primary comb line and the pump $\Delta f_{\rm pri}$, proportional to $(1/\beta_2)^{1/2}$, are observed at 2.36 THz, 3.25 THz and 7.17 THz, respectively. When $V_G = -1.8 \text{ V}$, the group velocity dispersion of the GMR becomes positive and hence it becomes harder to phase-match without local mode-crossing-induced dispersion. Figure 3b shows the optical spectra under carefully controlled laser-cavity detuning. In particular, at $V_G = -1 \text{ V}$, $\beta_2 \approx -62 \text{ fs}^2 \text{ mm}^{-1}$ and $\beta_3 \approx -9 \text{ fs}^3 \text{ mm}^{-1}$ the Kerr comb has a span of about 350 nm, with highly symmetrical shape. Interestingly, with $V_G = -1.2 \text{ V}$, $\beta_2 \approx -33 \text{ fs}^2 \text{ mm}^{-1}$ and $\beta_3 \approx -630 \, \text{fs}^3 \, \text{mm}^{-1}$, we observe a frequency comb spectrum spanning 600 nm, consistent with the general route of a smaller group velocity dispersion bringing about a broader comb spectrum. The comb spectrum is highly asymmetric, with the red-side comb line intensity contributions from Cherenkov radiation. The spectral peak of the Cherenkov radiation is determined by $3\beta_2/\beta_3$, matching the measurement results. In addition, such soliton perturbation and energy transfer can be used to stabilize the Kerr frequency comb¹³. When $V_G = -1.5 \text{ V}$, $\beta_2 \approx -8 \, \text{fs}^2 \, \text{mm}^{-1}$ while $\beta_3 \approx -213 \, \text{fs}^3 \, \text{mm}^{-1}$. Because β_2 here is small (less than 10 fs² mm⁻¹), it does not support a stable Kerr comb, the observed comb lines are not even, and the Cherenkov peak in the spectrum is also indistinguishable.

Figure 3c summarizes the gate tunability of the graphene Kerr combs, with $V_{\rm G}$ from 0 V to -2 V. For primary comb lines, their relative spectral location $\Delta f_{\rm pri} = |f_{\rm pri} - f_{\rm pump}|$ is strongly controlled, moving from 2.3 THz to 7.2 THz as $V_{\rm G}$ only changes from -1.0 V to -1.5 V. This modulation

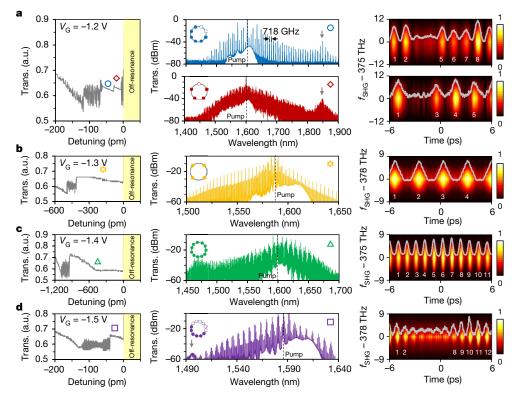


Fig. 4 | Soliton crystals of the gated graphene-nitride microresonator. $\bf a, d$, Soliton state with crystal-like defects including the single-soliton defect in $\bf a. b, c$, Periodic soliton crystal states with equally spaced soliton pulses. Panels $\bf a$ to $\bf d$ are achieved with gate voltages $V_{\bf G}$ tuned at different values, ranging from -1.2 V in $\bf a$ to -1.5 V in $\bf d$. Left panels: measured intensity transmission, illustrating the characteristic 'steps' associated

with soliton formation. Middle panels: corresponding optical spectra measurements. The pump locations are marked by black dashed lines and the Cherenkov radiation peaks are marked by grey arrows. Right panels: frequency-resolved second-harmonic autocorrelation maps of the soliton pulses. Here the grey curves show the real-time autocorrelation intensity traces.

is also influenced by the slight nonlinearity enhancement introduced by the graphene. For the full-span combs generated without missing comb lines, we also demonstrate electric-field control of their spectral span, from 38 THz to 82 THz with V_G from -1.0 V to -1.3 V (for this device, $|V_{\rm G}|$ > 1.3 V does not show a good coherent comb state). Moreover, we note that the gate-tuning changes the FSRs of the combs, from 89.6 GHz at -1.0 V to 89.9 GHz at -1.5 V. Such optoelectronic tunability enables different Kerr frequency combs with a variety of properties to exist in the same device. Figure 3d next illustrates the measured locations of the Cherenkov radiation peaks in comparison with the computed designs. In contrast to the primary comb lines, the third-order dispersion plays an important role in the Cherenkov radiation. We observed three Cherenkov peaks in the window from 1,400 nm to 2,000 nm, with spectral locations $\Delta f_c = |f_c - f_{\text{pump}}|$ at values of 26.3 THz ($V_G = -1.2 \text{ V}$), 49.2 THz ($V_G = -1.3 \text{ V}$) and 17.7 THz ($V_G = -1.5 \text{ V}$). The measured results well match the analytic calculation. In Fig. 3c and d, results are collected in the region of $-0.4\,\mathrm{V}$ to $-1.6\,\mathrm{V}$, because when V_G is more than -0.4 V, the Q factor of the GMR is too low for comb generation; and when V_G is less than -1.6 V, the group velocity dispersion is too small to ensure a stable comb.

We estimate the modulation speed of the GMR in Fig. 3e. With V_G tuning, the output comb line intensity within the filter window is modulated temporally. The modulation speed here is bounded by ion diffusion in the heterostructure, large ion-gel capacitance on the graphene, and the optical filter bandwidth. In our current proof-of-principle demonstration, to ensure that $|E_F|$ is sufficiently high, we use the ion-gel-based capacitor, the large capacitance $(7.2\,\mu F~cm^{-2})$ and slow ion diffusion (about $10^{-10}\,m^2\,s^{-1})$ of which limit the charge–discharge operation speed to less than hundreds of kilohertz. The optical filter bandwidth can be narrowed to improve the detection rate of the modulation by almost 7.5 times. In Fig. 3e, we show the modulated signal-to-noise ratio with a radiofrequency spectrum analyser, by

using optical filters with passband widths of 50 nm, 9 nm and 2 nm, respectively. Their corresponding bandwidths are 80 kHz, 200 kHz and 600 kHz (Extended Data Fig. 3). Although sub-megahertz modulation for the primary comb is successfully demonstrated, we note that fast modulation while preserving the full-grown Kerr comb across the entire modulation cycle could be much more challenging: with $V_{\rm G}$ tuning, not only the group velocity dispersion but also the FSR of the GMR is tuned. Compared with the primary combs shown in Fig. 3a, phase-matching of the full combs in Fig. 3b is much more sensitive: a slight variation in the FSR from the gate modulation may cause the Kerr comb to collapse. To achieve reliable, fast on–off switching in full-generated Kerr combs, inverse FSR compensation (for example, by temperature feedback) should be applied. Such sub-megahertz tunability for a Kerr comb could potentially be used in applications 31 such as precision measurements.

Dispersion is one of the most critical cavity parameters that defines the Kerr frequency comb dynamics. The broadband dispersion modulation controlled by the gate voltage of the graphene–nitride microresonator opens up the possibility of dynamically selecting the formation path of dissipative Kerr solitons and frequency combs. By using the gate-tunable GMRs, we can engineer the dispersion dynamically to form different soliton states through electrical control. With a fixed pump power of 2 W, Fig. 3f counts the soliton states achieved in measurements for gate voltages in the range $-1.6\,\mathrm{V}$ to $-1.1\,\mathrm{V}$, with the experimental conditions otherwise kept constant. In total, we have found soliton states with soliton numbers of 12, 11, 9, 8, 6, 5 and 4. More theoretical calculations and simulations are discussed in Supplementary Information section 1.4.

Figure 4 demonstrates four specific examples of soliton crystal states, under optimized gate voltages. Here the left panels show the measured intensity transmission, the middle panels demonstrate the optical spectra, and the right panels illustrate the frame-by-frame

frequency-resolved second-harmonic autocorrelation maps. These soliton states with low radiofrequency noise are achieved following Turing patterns and chaotic states before transition into the soliton states (Extended Data Fig. 4). This is characterized by a transmission step, by tuning the pump laser gradually into the cavity resonance. Figure 4a shows two examples of the soliton state with missing pulses, at a gate voltage of -1.2 V. The corresponding pump laser wavelength is around 1,600.2 nm. The optical spectra of these states are characterized by the apparent existence of groups of comb lines that are separated by multiple cavity FSRs. Within each comb group, weaker single-FSR comb lines are present, and they effectively connect all comb groups without any spectral gaps. For the examples shown in Fig. 4a and d, the comb groups are separated by 8 FSR, 5 FSR and 12 FSR respectively. In the time domain, the autocorrelation traces reveal the common features of missing pulses in the otherwise equally spaced soliton states with higher effective repetition rate. The self-organization of multiple soliton pulses into a train of equally spaced pulses resembles the crystallization process and is therefore termed a soliton crystal³², and the missing pulse structure is analogous to defects in crystal lattices. Our graphene-nitride heterogeneous microresonator thus provides a platform for study of soliton physics that is tunable through the gate-voltage and Fermi level. We also note that when the soliton crystals are formed, the emitted soliton Cherenkov radiations are sharp and narrow, as marked by the grey arrows in Fig. 4.

Soliton crystals are formed because of the strong mode interaction and intracavity interferences, and thus their evolution dynamics depend critically on the exact dispersion profile of the microresonator. By further optimizing the group velocity dispersion and third-order dispersion through gate tuning, we demonstrate two periodic soliton crystal states. Figure 4b shows a four-soliton state with $V_G = -1.3 \,\mathrm{V}$ and pump laser at approximately 1,584.2 nm, while Fig. 4c shows a 11-soliton state with $V_G = -1.4 \,\mathrm{V}$ and pump laser at approximately 1,600.1 nm. Intriguingly, these soliton crystal states show remarkable stability, and they can robustly survive a pump power fluctuation up to ± 2 dB, or wavelength offset up to ± 300 pm. The soliton crystal formation is also akin to harmonic mode-locking in which a stable highrepetition-rate pulse train can be attained even in longer cavities, and it is of interest in applications such as high-speed communication, comb spectroscopy and data storage. This realization of a charge-tunable graphene heterostructure for controllable frequency combs and soliton dynamics opens a new architecture at the interface of single-atomiclayer nanoscience and ultrafast optoelectronics.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0216-x.

Received: 7 February 2017; Accepted: 19 March 2018; Published online 11 June 2018.

- Udem, T., Holzwarth, R. & Hansch, T. Optical frequency metrology. Nature 416, 1. 233-237 (2002).
- Kippenberg, T., Holzwarth, R. & Diddams, S. Microresonator-based optical frequency combs. Science 332, 555-559 (2011).
- Cingöz, A. et al. Direct frequency comb spectroscopy in the extreme ultraviolet.
- Nature **482**, 68–71 (2012). Ideguchi, T. et al. Coherent Raman spectro-imaging using laser frequency combs. Nature 502, 355-358 (2013).
- Steinmetz, T. et al. Laser frequency combs for astronomical observations. Science 321, 1335–1337 (2008).
- Huang, S.-W. et al. Mode-locked ultrashort pulse generation from on-chip normal dispersion microresonators. Phys. Rev. Lett. **114**, 053901 (2015).
- Saglamyurek, E. et al. Broadband waveguide quantum memory for entangled photons. Nature 469, 512-515 (2011).
- Del'Haye, P. et al. Optical frequency comb generation from a monolithic microresonator. Nature 450, 1214-1217 (2007).
- Moss, D. J., Morandotti, R., Gaeta, A. L. & Lipson, M. New CMOS-compatible platforms based on silicon nitride and Hydex for nonlinear optics. Nat. Photon.
- Yang, Q. F., Yi, X., Yang, K. Y. & Vahala, K. Stokes solitons in optical microcavities. Nat. Phys. 13, 53-57 (2016).

- 11. Xue, X. et al. Mode-locked dark pulse Kerr combs in normal-dispersion microresonators. Nat. Photon. 9, 594-600 (2015).
- Huang, S.-W. et al. A broadband chip-scale optical frequency synthesizer at 2.7×10^{-16} relative uncertainty. *Sci. Adv.* **2**, e1501489 (2016).
- 13. Brasch, V. et al. Photonic chip-based optical frequency comb using soliton Cherenkov radiation. Science 351, 357–360 (2016).
- Marin-Palomo, P. et al. Microresonator-based solitons for massively parallel coherent optical communications. Nature 546, 274-279 (2017).
- 15. Del'Haye, P. et al. Phase-coherent microwave-to-optical link with a selfreferenced microcomb. Nat. Photon. 10, 516-520 (2016).
- 16. Herr, T., Brasch, V., Jost, J., Wang, C. & Kondratiev, N. Temporal solitons in optical microresonators. Nat. Photon. 8, 145-152 (2014).
- Wang, F. et al. Gate-variable optical transitions in graphene. Science 320, 206-209 (2008).
- 18. Li, Z., Henriksen, E., Jiang, Z., Hao, Z. & Martin, M. Dirac charge dynamics in graphene by infrared spectroscopy. Nat. Phys. 4, 532-535 (2008).
- Bonaccorso, F., Sun, Z., Hasan, T. & Ferrari, A. Graphene photonics and optoelectronics. Nat. Photon. 4, 611-622 (2010).
- Vakil, A. & Engheta, N. Transformation optics using graphene. Science 332, 1291-1294 (2011).
- 21. Gu, T. et al. Regenerative oscillation and four-wave mixing in graphene optoelectronics. Nat. Photon. 6, 554-559 (2012).
- Liu, M. et al. A graphene-based broadband optical modulator. Nature 474,
- 64–67 (2011). Phare, C., Lee, Y., Cardenas, J. & Lipson, M. Graphene electro-optic modulator with 30 GHz bandwidth. Nat. Photon. 9, 511-514 (2015).
- Koppens, F. et al. Photodetectors based on graphene, other two-dimensional materials and hybrid systems. Nat. Nanotech. 9, 780-793 (2014).
- Grigorenko, A., Polini, M. & Novoselov, K. Graphene plasmonics. Nat. Photon. 6, 749–758 (2012).
- Chakraborty, S. et al. Gain modulation by graphene plasmons in aperiodic lattice lasers. Science 351, 246-248 (2016).
- Xu, Y. et al. Holey graphene frameworks for highly efficient capacitive energy storage. Nat. Commun. 5, 4554 (2014).
- 28. Das, A. et al. Monitoring dopants by Raman scattering in an electrochemically top-gated graphene transistor. Nat. Nanotech. 3, 210-215 (2008).
- Javerzac-Galy, C. et al. Excitonic emission of monolayer semiconductors near-field coupled to high-Q microresonators. Nano Lett. 18, 3138-3146 (2018).
- 30. Sorianello, V. et al. Graphene phase modulator. Nat. Photon. 12, 40-44 (2018).
- Huang, S. et al. Globally stable microresonator Turing pattern formation for coherent high-power THz radiation on-chip. Phys. Rev. X 7, 041002 (2017).
- 32. Cole, D., Lamb, E., Del'Haye, P., Diddams, S. A. & Papp, S. B. Soliton crystals in Kerr resonators. Nat. Photon. 11, 671-676 (2017).

Acknowledgements We thank J. Yang, B. Li, T. Itoh, H. Liu and X. Xie for discussions. Graphene fabrication was supported by the Nanoelectronics Research Facilities (NRF) of UCLA. The authors acknowledge support from the National Science Foundation (NSF; DMR-1611598, CBET-1520949 and EFRI-1741707), the University of California National Laboratory research program (LFRP-17-477237), the Office of Naval Research (N00014-16-1-2094) and the Air Force Office of Scientific Research (FA9550-15-1-0081). X.F.D. acknowledges support from the Office of Naval Research (N00014-15-1-2368) and Y.H. acknowledges support from the NSF (EFRI-1433541). This work is also supported by the National Science Foundation of China (61705032) and the 111 project of China (B14039).

Reviewer information Nature thanks T. Tanabe and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.Y. and S.-W.H. designed and led the work on the graphene-nitride frequency combs including the first and detailed measurements, the gate-tuned ultrafast optics measurements and the numerical designs. Y.L. and Z.Y.F. performed the graphene-nitride integration, conducted relevant electrical measurements and device optimizations. B.Y., C.C., S.-W.H., M.H., M.Y. and D.-L.K. performed silicon nitride chip and device processing. Y.H. and X.F.D. supervised graphene material preparation, device fabrication and electronic measurements. B.Y., A.K.V., S.-W.H. and C.W.W. performed the measured data analysis, on the frequency comb, radiofrequency and ultrafast correlation measurements. S.-W.H., B.Y. and Y.N.L. provided the theory and numerical calculations. All authors discussed the results. B.Y., S.-W.H., Y.L., Y.J.R. and C.W.W. prepared the manuscript. C.W.W. led and supported this research.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0216-x

Supplementary information is available for this paper at https://doi. org/10.1038/s41586-018-0216-x

Reprints and permissions information is available at http://www.nature.com/

Correspondence and requests for materials should be addressed to B.Y., S.-W.H., X.F.D. or C.W.W.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

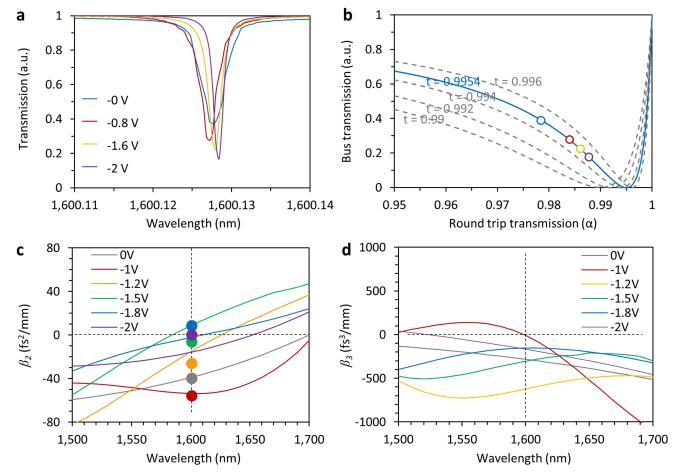
Theoretical analysis. The refractive index of graphene, $n_{\rm G}$, is determined by its permittivity $\varepsilon_{\rm G}$ as $n_{\rm G}=\varepsilon_{\rm G}^{1/2}$, where $\varepsilon_{\rm G}=\{-{\rm Im}(\sigma_{\rm G})+i{\rm Re}(\sigma_{\rm G})\}/\{2\pi f\Delta\}$ (ref. 20); here $\sigma_{\rm G}$ is the conductivity of graphene, f is the optical frequency, and $\Delta=0.4\,{\rm nm}$ is the thickness of the graphene monolayer. Particularly, $\partial^n{\rm Re}(n_{\rm G})/\partial\lambda^n$ determines the nth-order dispersion, while ${\rm Im}(n_{\rm G})$ determines the waveguide loss. In Supplementary Information section 1.1, we describe in more detail how the transmission of graphene is determined by its quasi Fermi level $E_{\rm F}$. By gating graphene via an external field, one can conveniently control both the group velocity dispersion β_2 and the third-order dispersion β_3 of a graphene monolayer. Kerr comb generation in the time domain is governed by the well-known Lugiato–Lefever equation in the GMR. In Supplementary Information section 1.2, we analyse the comb formation dynamics by LLE modelling. In Supplementary Information section 1.3, we describe the third-order nonlinearity of graphene. In Supplementary Information section 1.4, we provide detailed simulations of the soliton generations in the GMR

Device design and fabrication. First, in a silicon foundry, we nanofabricated a high-Q silicon nitride microresonator with measured loaded $Q \approx 1.6 \times 10^6$ (intrinsic $Q \approx 1.8 \times 10^6$) and FSR ≈ 90 GHz in a 350- μ m-diameter ring structure. The nitride core has a $1,200 \times 800 \,\mathrm{nm^2}$ cross-section, a 600 nm gap to the input-output coupling waveguide of $1,000 \times 800 \, \text{nm}^2$ cross-section, and a top oxide cladding. Next, single-atomic-layer graphene was grown using a chemical vapour deposition method and transferred onto the exposed region of the nitride ring (with etched SiO₂ window). The monolayer graphene was then lithographically patterned and oxygen plasma etched into an $80 \,\mu\text{m} \times 100 \,\mu\text{m}$ sheet. Metallization of source– drain electrodes was achieved through standard photolithography, followed by electron-beam evaporation of Ti/Au (20/50 nm thick). Here the electrode pad size was $80 \mu m \times 60 \mu m$. Subsequently, we integrated an ionic liquid (DEME-TFSI, N,Ndiethyl-N-methyl-N-(2-methoxyethyl)ammonium bis(trifluoromethanesulfonyl)imide) as the gate dielectric, resulting in an electric double-layer graphene transistor. More details are shown in Supplementary Information section 2 and Supplementary Fig. 12.

Experimental set-ups. We implemented a temperature-controlled optical set-up for the frequency comb generation. The spectral tunable range of our drive laser is 1,480 nm to 1,640 nm, and the maximum output power of our erbium-doped fibre amplifier (EDFA, BKtel) in the L-band is 3.16 W (35 dBm). The GMR transmission is measured by using the same tunable laser, swept through its full wavelength tuning range at a speed of 40 nm s^{-1} , to obtain dispersion and Q factors. A fibre-coupled hydrogen cyanide gas cell (HCN-13-100, Wavelength References Inc.) and an unbalanced fibre Mach-Zehnder interferometer are used for calibration. To measure the stability and soliton states of our frequency comb, heterodyne and autocorrelation measurements are implemented. For the heterodyne measurement, a stable continuous-wave laser with narrow linewidth (300 kHz, New Focus) is applied as the heterodyne reference for the beat notes. For the autocorrelation measurement, a fibre with zero group velocity dispersion, made of a 7-m dispersion-compensating fibre and a 15-m single-mode fibre, is used to guide the microresonator output to the autocorrelation set-up with minimal pulse broadening and distortion. More details about the experimental set-ups are shown in Supplementary Figs. 13-16.

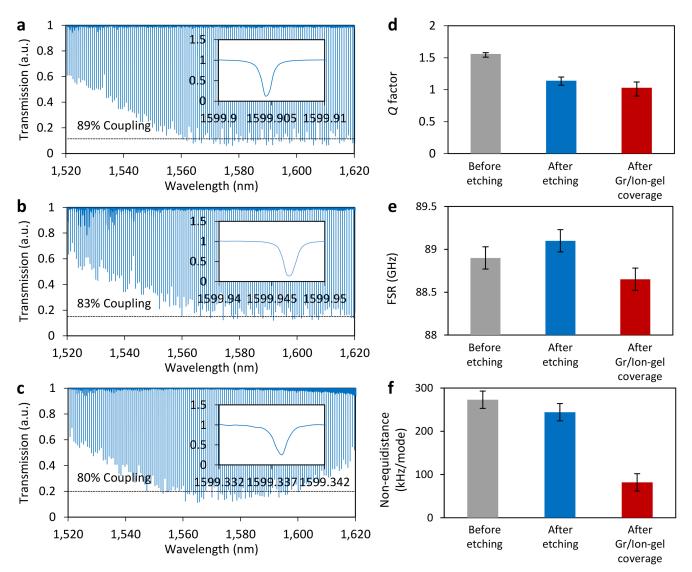
Soliton-step evolution process in the graphene frequency comb. Complementary to Fig. 4, Extended Data Fig. 4 illustrates the soliton states at a gate voltage of $-1.2\,\mathrm{V}$, for different laser-cavity detunings. With the simultaneous optical and radiofrequency spectra measurements, the frequency comb initiates from the Turing pattern (state i) into the high-noise patterns (states ii and iii, with sub-comb competition) before settling down into the low-noise soliton comb (states iv and v). With further detuning, the soliton comb goes back into the high-noise regime (state vi), owing to the thermal instability in the cavity. We then beat the comb lines of the soliton states with a continuous-wave reference laser with two examples for the eight-soliton state (state iv) and the four-soliton state (state v). Here, the 9th mode and the 56th mode denote the offsets from the pump line. The beat notes show an intensity contrast ratio of more than 40 dB, with a linewidth of 200 kHz, verifying that there is clean and stable soliton generation.

Data availability. The data that support the findings of this study are available from the corresponding authors on reasonable request.



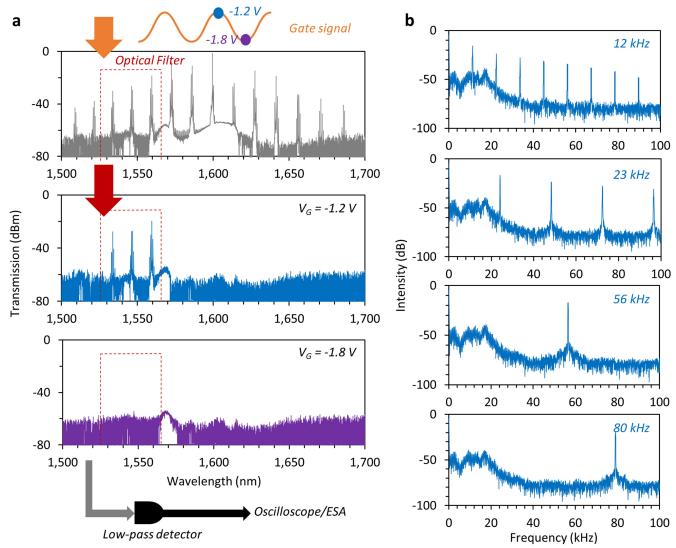
Extended Data Fig. 1 | **Measured gate-tunable coupling and dispersions in a GMR. a**, Dips at approximately 1,600 nm, with different $V_{\rm G}$. **b**, Correlation of the round-trip transmissions and the bus transmissions for the resonator, obeying $T=(\alpha-|t|)^2/(\alpha-\alpha|t|)^2$. Here, $1-\alpha$ is the cavity loss per round trip, and 1-t is the bus-to-cavity coupling rate. In our

experiment, the graphene ring resonator is under-coupled originally, as the blue dot shows. \mathbf{c} , Group velocity dispersion in range of 1,500 nm to 1,700 nm. Here, the curves show the calculated results, while dots show measured data. \mathbf{d} , Calculated third-order dispersion in range of 1,500 nm to 1,700 nm.



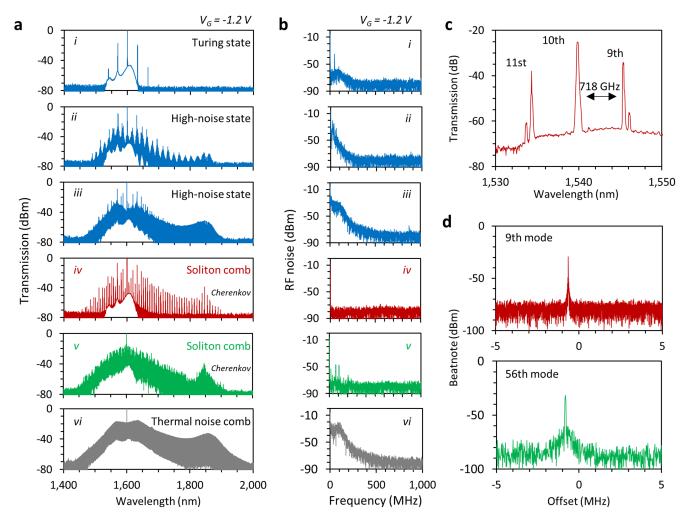
Extended Data Fig. 2 | Comparative optical transmissions of the heterogeneous graphene–nitride ring. a, Spectral transmission of the silicon nitride ring resonator under the silica overcladding. b, Spectral transmission of the silicon nitride ring resonator after buffer-oxide etching to remove the silica overcladding. c, Spectral transmission of the graphene/

ion-gel-based nitride ring resonator, heavily p-doped ($V_{\rm G}=-2$ V). d, Loaded Q factor around 1,600 nm. e, FSR, which is sensitive to the geometry modification. f, Mode non-equidistances, D_2 . d and e are measured at $\lambda=1,600$ nm. In this figure, the error bars denote the typical system error.



Extended Data Fig. 3 | An implementation of the graphene primary frequency comb gate-modulation. a, Method for measuring the modulated comb. Keeping bias $V_{\rm G}=-1.2$ V, we control the laser-cavity detuning to generate a primary comb such as the grey spectrum shown here. To filter off the 1,600-nm continuous-wave pump, we apply a C-band filter, selecting the comb lines in the C-band only. A signal generator (maximum amplitude of 2 V, HP3312) is applied to modulate the gate

voltage between $-1.2\,\mathrm{V}$ and $-1.8\,\mathrm{V}$. In this process, primary comb lines in the filter window are modulated by the gate signal; the modulation is monitored by using an oscilloscope (500 MHz, Rigol DS1054) and an electrical spectrum analyser (ESA, 3 GHz, Agilent CXA9000A). **b**, Examples of radiofrequency spectra of the modulated combs, filtered by an optical filter (1,530 nm to 1,570 nm).



Extended Data Fig. 4 | Example measurements of the graphene soliton comb formation. a, Under $V_{\rm G}=-1.2$ V (Fermi level 0.59 eV), when the wavelength of the pump ($\lambda_{\rm p}$) is tuned from 1,600.00 nm to 1,600.23 nm, the Kerr frequency comb is generated gradually. When $\lambda_{\rm p}$ is tuned between 1,600.15 nm and 1,600.19 nm, two multi-soliton states with low phase noise are achieved (states iv and v). b, Corresponding radiofrequency (RF) amplitude noise of the six states. In a and b, the pump power is kept at

34.5 dBm. Cherenkov radiation of the multi-soliton comb is narrow and sharp. **c**, Zoom-in of the eight-soliton crystal spectrum. The FSR changes from 89 GHz to 718 GHz, owing to the soliton-crystal-based longitude mode interaction. **d**, Beat note for the comb lines of the eight-soliton state (red; ninth comb line offset from the pump) and the four-soliton state (green; 56th comb line offset from the pump).



Comprehensive suppression of single-molecule conductance using destructive σ -interference

Marc H. Garner^{1,9}, Haixing Li^{2,6,9}, Yan Chen^{3,9}, Timothy A. Su^{4,7}, Zhichun Shangguan^{3,8}, Daniel W. Paley^{4,5}, Taifeng Liu³, Fay Ng⁴, Hexing Li³, Shengxiong Xiao^{3*}, Colin Nuckolls^{3,4*}, Latha Venkataraman^{2,4*} & Gemma C. Solomon^{1*}

The tunnelling of electrons through molecules (and through any nanoscale insulating and dielectric material¹) shows exponential attenuation with increasing length², a length dependence that is reflected in the ability of the electrons to carry an electrical current. It was recently demonstrated³⁻⁵ that coherent tunnelling through a molecular junction can also be suppressed by destructive quantum interference⁶, a mechanism that is not length-dependent. For the carbon-based molecules studied previously, cancelling all transmission channels would involve the suppression of contributions to the current from both the π -orbital and σ -orbital systems. Previous reports of destructive interference have demonstrated a decrease in transmission only through the π -channel. Here we report a saturated silicon-based molecule with a functionalized bicyclo[2.2.2]octasilane moiety that exhibits destructive quantum interference in its σ -system. Although molecular silicon typically forms conducting wires⁷, we use a combination of conductance measurements and ab initio calculations to show that destructive σ -interference, achieved here by locking the silicon-silicon bonds into eclipsed conformations within a bicyclic molecular framework, can yield extremely insulating molecules less than a nanometre in length. Our molecules also exhibit an unusually high thermopower (0.97 millivolts per kelvin), which is a further experimental signature of the suppression of all tunnelling paths by destructive interference: calculations indicate that the central bicyclo[2.2.2]octasilane unit is rendered less conductive than the empty space it occupies. The molecular design presented here provides a proof-of-concept for a quantum-interference-based approach to single-molecule insulators.

In molecular electronics, the focus has primarily been on creating devices in which single molecules, bridging the gap between two metal electrodes, mimic the functionality of classical electronic components such as resistors, diodes or switches⁸. Development of highly insulating molecules has been neglected, in part because the coherent electron transport mechanism becomes exponentially more efficient as the dimensions are scaled down. Designing highly insulating sub-nanometre molecules is therefore difficult. A breakthrough at the single-molecule level might hint at a strategy for overcoming the fundamental challenge of direct tunnelling through insulators in classical devices^{9,10}. Here we propose a strategy to create an ideal single-molecule insulator: a molecule bridging a nanometre gap between two metal electrodes across which electronic transmission is completely suppressed. We propose that an ultimate molecular insulator can be achieved using molecules that exhibit complete destructive interference in the transmission.

Figure 1a illustrates the classical example of electronic tunnelling across a vacuum potential barrier^{11,12}. The electronic wavefunction,

which extends into the classical energetically forbidden region between the electrodes, is attenuated; the extent of the attenuation depends exponentially on the separation between the electrodes as shown in Fig. 1b. Inserting a molecule or material between the electrodes will generally result in increased electronic transmission across the gap. Figure 1c illustrates a molecular junction in which transmission is mediated by the frontier molecular orbitals that are coupled to the metal electrodes. As long as the orbitals are not energetically close to the Fermi level of the metal, transmission is in an off-resonant regime, as shown in Fig. 1d. In this regime, if the length of the molecule is increased by adding identical repeat units, such as a methylene group in a linear alkane, transmission decreases exponentially with increasing length (inset, Fig. 1d)². Thus, although alkanes^{8,13} and molecular siloxane¹⁴ have good insulating properties, the transmission across these metalmolecule-metal junctions can always be lowered by removing the molecule. As the junction length decreases, conductance is not effectively suppressed in these insulating systems as the transmission increases exponentially.

Here, we demonstrate a new strategy that relies on a destructive interference effect to suppress coherent transmission across a molecular junction. This effect is a consequence of phase-coherence of the transiting electrons and cannot be modelled by a simple tunnel barrier¹¹. In theory, interference can lead to complete cancellation of the tunnelling probability; consequently, a molecule with complete suppression of the transmission will be less conducting than a vacuum gap of the same dimensions. Destructive interference effects have been demonstrated extensively in recent years, where the focus has been on carbon-based π -conjugated molecules such as a meta-linked benzene molecule, as illustrated in Fig. 1e 4,8. In such carbon-based wires, transmission can occur through both a π -channel and a σ -channel. Destructive interference annuls transmission completely only in the π -channel and leaves the σ -channel unchanged, as illustrated schematically in Fig. 1f 13 . Thus, the π -conjugated carbon-based systems studied previously cannot achieve complete suppression of the transmission. The lower limit of the conductance will always be set by the conductance of the σ -channel and comparable to an alkane of similar length¹⁵. Additionally, through-space injection from the electrodes into conducting paths can short-circuit the interference effect¹⁶. Therefore, we turn to saturated silicon-based molecules that have solely a σ -orbital system.

In contrast to alkanes, silanes have strong σ -conjugation through their backbones^{17–19}, and linear silane wires are consequently good electrical conductors⁷. It has previously been suggested that destructive quantum interference can be seen in alkanes and silanes in conformations with small dihedral angles in the molecular backbone^{20,21}. We thus designed bicyclo[2.2.2]octasilane, which has *cisoid* Si–Si–Si–Si dihedral angles ranging from 15° to 20° for its three silane bridges. We explore the insulating properties of two methylthiomethyl-functionalized

¹Nano-Science Center and Department of Chemistry, University of Copenhagen, Copenhagen, Denmark. ²Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY, USA. ³The Education Ministry Key Lab of Resource Chemistry, Shanghai Key Laboratory of Rare Earth Functional Materials, Optoelectronic Nano Materials and Devices Institute, Department of Chemistry, Shanghai Normal University, Shanghai, China. ⁴Department of Chemistry, Columbia University, New York, NY, USA. ⁵Columbia Nano Initiative, Columbia University, New York, NY, USA. ⁶Present address: Department of Chemistry, Columbia University, New York, NY, USA. ⁷Present address: Department of Chemistry, University of California, Berkeley, Berkeley, CA, USA. ⁸Present address: School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, Shanghai, China. ⁹These authors contributed equally: M. H. Garner, H. Li, Y. Chen. *e-mail: xiaosx@shnu. edu.cn; cn37@columbia.edu; lv2117@columbia.edu; gsolomon@chem.ku.dk

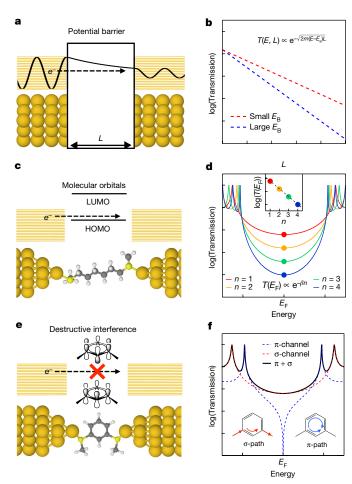


Fig. 1 | Schematic illustration of coherent electron transport and model transmission. a, b, Tunnelling through a simple potential energy barrier, E_B , where the transmission probability, T(E), decays exponentially with length, L. c, d, Schematic of coherent transmission across a molecule. The transmission decays exponentially across a molecular series with n repeat units (inset)². LUMO, lowest unoccupied molecular orbital; HOMO, highest occupied molecular orbital e, f, Schematic of coherent transmission across a molecule for which there is destructive quantum interference between contributions from different π -orbitals, as is the case for a meta-coupled benzene unit³². The transmission has contributions from σ - and π -channels; the π -channel is fully suppressed at the antiresonance, but, owing to the σ -contribution, the total transmission does not have an antiresonance¹³.

bicyclo[2.2.2]octasilanes, **Si222** and **Si2-Si222-Si2**, and compare them with the linear silane counterpart **Si4**, which has the same number of silicon atoms as the shortest single path through **Si222** (structures shown in Fig. 2a). Bicyclo[2.2.2]octasilanes have previously been studied theoretically with alternative binding groups, but in that case no appreciable σ -interference was observed²².

We first calculate the Landauer transmission for Au-molecule—Au junctions, using density functional theory (DFT) as detailed in the Methods section, and present the results in Fig. 2b. The transmission at the Fermi energy is two orders of magnitude lower for Si222 than for its linear counterpart Si4, owing to a sharp antiresonance in the transmission close to the Fermi energy. The extended molecule, Si2-Si222-Si2, also has an antiresonance close to the Fermi energy, indicating a clear suppression of the transmission by destructive quantum interference. These results are for one of the three conformations of Si222; others are shown in Extended Data Fig. 1.

To probe the origin of the interference, we resolve the interatomic transmission pathways²⁰ as illustrated in Fig. 2c–e. The transmission through **Si4** (Fig. 2c) shows one dominant through-bond pathway with no sign of any interference features. The transmissions through **Si222**

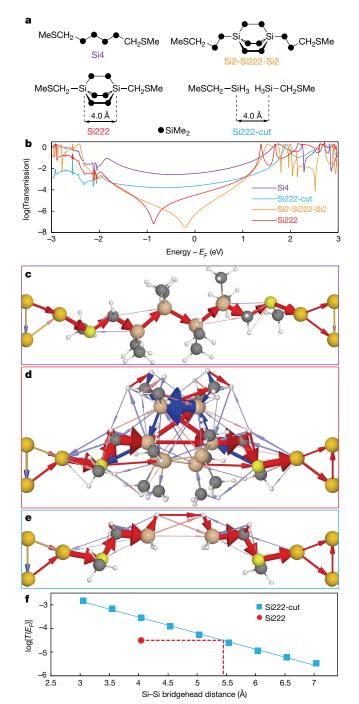


Fig. 2 | Calculated transport properties of Si4, Si222, Si222-cut and Si2-Si222-Si2. a, Structures of the molecules investigated. b, Landauer transmission plotted semi-logarithmically against energy relative to the Fermi energy. c-e, Interatomic transmission pathway analysis of (c) Si4, (d) Si222 and (e) Si222-cut. The size of the arrows between atoms is proportional to the magnitude of the transmission contribution, and the colour of the arrows represents the direction (the sign). Si222 exhibits lower levels of transmission than Si222-cut across a wide energy range, indicating that removing the chemical bonds that link the bridgehead silicon atoms increases the transmission. f, Transmission at the Fermi energy plotted against bridgehead distance of Si222 and Si222-cut. Solid blue line is a linear fit to the data of Si222-cut; dashed red line is a visual guide for comparison.

and Si2-Si222-Si2 (Fig. 2d and Extended Data Fig. 2c) exhibit ringcurrent reversal in the bicyclic structure, as indicated by the red and blue arrows, where the sign (colour) reverses in the energy range around the antiresonance (compare with Extended Data Fig. 2c-f). This is a clear signature of quantum interference; similar features were seen in π -conjugated molecules²⁰. Removing through-bond paths by cutting away the silicon bridges one at a time while passivating the remaining silicon atoms with hydrogens reveals that the interference is gradually lifted as through-bond pathways are disrupted and only the through-space paths remain, as shown in Extended Data Fig. 3. This analysis indicates that the cancellation may occur from destructive interference between the through-space and through-bond paths in Si222.

As the origin of interference is in the bicyclic moiety of Si222, we compare this unit with the space that it occupies. With all three bridges cut off (Si222-cut), the transmission (Fig. 2b) at the Fermi energy is an order of magnitude higher than for Si222. That is, the electronic structure of the full molecule is more effective in suppressing transmission than the decay of the wavefunction in the gap between the bridgehead silicon atoms (that is, the terminal atoms of the bicyclic unit) in Si222cut. The pathways of Si222-cut (Fig. 2e) show that as the through-bond pathways have been removed, only through-space pathways remain. For comparison, if we perform the same operation with Si4 and remove a Si(CH₃)₂ unit, the transmission simply drops as shown in Extended Data Fig. 4. Finally, in Fig. 2f we plot the transmission at the Fermi energy as we manipulate the size of the vacuum gap between the two disjointed silyl groups in Si222-cut. This analysis reveals that the transmission in Si222 corresponds to that of a vacuum gap of just over 5.4 Å, considerably larger than the bridgehead distance of 4.0 Å, and conclusively predicts that destructive interference enables the bicyclic moiety to function as an extremely insulating molecular unit—more insulating than can be achieved by a gap of the same dimensions.

We synthesize **Si4**, **Si222** and **Si2-Si222-Si2** to probe their insulating properties. Briefly, we react dodecamethylbicyclo-[2.2.2] octasilanyl-1,4-dianion²³ with chloromethyl methyl sulfide and 1-chloro(2-methylthiomethyl)tetramethyldisilane to obtain **Si222** and **Si2-Si222-Si2**, respectively, as shown in the scheme in Fig. 3a. **Si4** was synthesized by previously reported methods⁷.

We measure the single Au-molecule–Au junction conductance using a scanning tunnelling microscope break junction (STM-BJ) technique as detailed in the Methods section^{24,25}. In Fig. 3b, we plot logarithmically binned 1D conductance histograms compiled from about 10,000 to 30,000 conductance traces for Si4, Si222 and Si2-Si222-Si2, where we observe clear molecular conductance peaks. Two-dimensional conductance–displacement data are shown in Extended Data Fig. 5. In good qualitative agreement with the theoretical predictions presented in Fig. 2, the conductance of Si222 is an order of magnitude lower than that of Si4, and the conductance of Si2-Si222-Si2 is an order of magnitude lower than that of Si222.

We now turn to thermopower measurements²⁶ to probe the slope of the transmission function close to the Fermi energy; if this is large in magnitude, it is highly indicative of an interference effect across the junction^{27,28}. Measurements are carried out as detailed in the Methods section²⁹. For Si2-Si222-Si2, we measure a thermopower of about $0.97 \pm 0.03 \,\mathrm{mV K^{-1}}$ (Fig. 3c), larger than the highest previously reported thermopower of approximately $-33 \mu V K^{-1}$ for C_{60} dimer³⁰ and more than 30 times that of its linear silane counterpart Si8 at about $35 \pm 17 \,\mu\text{V K}^{-1}$ (Extended Data Fig. 6). The slope of the calculated transmission is smaller than the experimental values. However, this result is extremely sensitive to the position of the antiresonance relative to the Fermi energy (Extended Data Fig. 6). Given the known errors inherent to DFT calculations and the impact of the experimental environment on the relative alignment of the Fermi level, reliable quantitative agreement between theory and experiment is not achieved for the thermopower^{29,31}. Nevertheless, the large experimental thermopower coefficient presents evidence of a sharp antiresonance in the transmission function near the Fermi level due to destructive quantum interference in this σ -coupled bicyclo[2.2.2]octasilane moiety.

We find, however, that the thermopower for **Si222** is below our measurement resolution, possibly owing to its bulky nature which makes it susceptible to through-space injection directly into

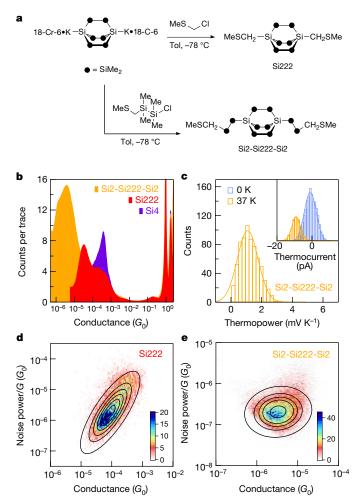


Fig. 3 | Synthesis scheme and experimental single-molecule conductance, thermopower and noise data. a, Si222 and Si2-Si222-Si2 were synthesized in 40% and 47% yields, respectively, following the scheme shown. Tol, toluene; 18-Cr-6, crown ether 1,4,7,10,13,16-hexaoxacyclooctadecane. **b**, Logarithmically binned 1D conductance histograms for Si222, Si4 and Si2-Si222-Si2 (100 bins per decade). The bias used for these measurements is: 220 mV for Si222, 45 mV for Si4 and 750 mV for Si2-Si222-Si2. The noise floors for the Si222 and Si4 histograms have been removed from this figure as they overlap with the peak of Si2-Si222-Si2. Note that the conductance for Si2-Si222-Si2 measured here is an upper bound owing to the large bias used in the measurements. Conductance quantum $G_0 = 2e^2/h$, where eis the charge on an electron, and h is Planck's constant. c, Histogram of thermopower determined from the zero-bias thermoelectric current and junction conductance for each Si2-Si222-Si2 junction measured at a temperature difference of 37 K (approximately 700 junctions). Inset: Histograms of thermoelectric current at temperature differences of 37 K and 0 K between the tip and the substrate. The orange and blue curves are Gaussian fits to the distributions. d, e, Two-dimensional histograms of normalized flicker noise power plotted against average junction conductance for: (d) Si222 (10,425 traces) and (e) Si2-Si222-Si2 (3,000 traces). We determine, from these data, that the noise power scales as G^2 for Si222 and $G^{0.9}$ for Si2-Si222-Si2.

the bicyclic moiety, as detailed in Extended Data Fig. 2a and b. Such a direct injection will result in a smaller slope of the transmission function near the Fermi level, thus yielding a low thermopower with only a modest increase in the conductance. If many of the junctions measured are not fully extended, the thermopower will be low despite the low conductance.

Motivated by the low thermopower of Si222 but high thermopower in Si2-Si222-Si2, we quantify the electron injection path into these molecules through their flicker noise characteristics. Flicker noise

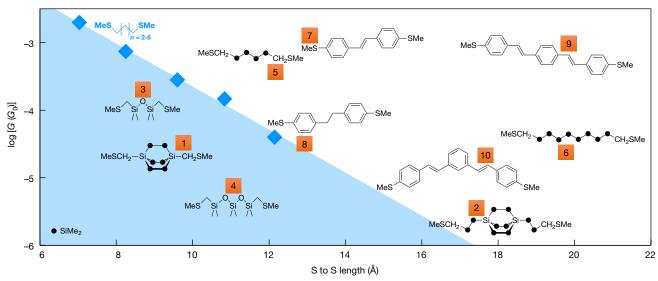


Fig. 4 | Experimental single Au-molecule-Au junction conductance against molecular length. The experimental conductance is plotted semilogarithmically against theoretical direct sulfur-to-sulfur length of the molecules. The blue area denotes molecules that are below the

conductance decay line of alkanes. All conductance values are determined from Au–molecule–Au junctions in published results^{7,14,33} and Extended Data Fig. 8a.

power scales differently with conductance for through-bond and through-space injection from the electrode into the molecule 16. We determine the correlation between the flicker noise power and the respective junction conductance as detailed in the Methods section. By plotting 2D histograms of normalized flicker noise power versus junction conductance G (Fig. 3d and e), we find that the noise power scales as G^2 for Si222 and as $G^{0.9}$ for Si2-Si222-Si2. A scaling of G^2 indicates that conductance is mediated by a through-space coupling, probably from the electrodes directly into the bicyclic structure. The considerable through-space injection in Si222 is consistent with the high-conductance shoulder seen in the conductance histograms (see Fig. 3b and Extended Data Fig. 5b). By contrast, a scaling of $G^{0.9}$ corresponds to through-bond-dominated transport for the longer Si2-Si222-Si2 molecule, as the bicyclic moiety is better protected from the electrodes. Control measurements for **Si4** (shown in Extended Data Fig. 5d) show a through-bond conduction as can be expected from the theoretical results in Fig. 2c.

To put the results into the context of highly insulating molecules, we compare the experimental conductance of a range of thiomethyl functionalized molecules with that of Si222 in Fig. 4. We propose that conductance and length together can provide a measure of the effectiveness of a molecular insulator (analogous to resistivity): thus, we plot the experimental conductance against the calculated sulfur-tosulfur distance. As alkanes have historically set the standard for insulating molecular moieties, the shaded blue region shows the molecular structures for insulators that are better than alkanes. Si222 is below the alkane line and even falls just below the strongly insulating siloxane molecular wires ¹⁴ (Extended Data Fig. 7a). Those π -conjugated molecules with destructive interference have much lower conductances than those that do not have destructive interference, in agreement with previous studies. However, their experimentally determined conductance (for example, molecule 10) indicates that they are less insulating than alkanes of the same length. The trends presented in Fig. 4 are independent of binding groups, as measured conductance of amine-terminated molecules display the same trend (Extended Data Fig. 7b).

In recent work, we have shown that linear silanes and siloxanes, to good approximation, mimic the properties of their widely used bulk materials 14 . Here we report a different bicyclic molecular form of silicon that is superior to siloxane as a single-molecule insulator. In contrast to previously reported insulating molecules, the mechanism is due to a destructive σ -interference effect within the bicyclic structure

where the Si atoms are constrained to small dihedral angles. Thus, bicyclic silanes represent saturated molecules with clear experimental and calculated signatures of destructive σ -interference. The realization of σ -interference enables the rational design of a new class of highly insulating molecules with extremely low conductance relative to their length, and with very high thermopower.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0197-9.

Received: 14 September 2017; Accepted: 15 March 2018; Published online 6 June 2018.

- Kingon, A. I., Maria, J.-P. & Streiffer, S. K. Alternative dielectrics to silicon dioxide for memory and logic devices. *Nature* 406, 1032–1038 (2000).
- Nitzan, A. Électron transmission through molecules and molecular interfaces. Annu. Rev. Phys. Chem. 52, 681–750 (2001).
- Mayor, M. et al. Electric current through a molecular rod—relevance of the position of the anchor groups. Angew. Chem. Int. Ed. 42, 5834–5838 (2003).
- Guédon, C. M. et al. Observation of quantum interference in molecular charge transport. Nat. Nanotech. 7, 305–309 (2012).
- Arroyo, C. R. et al. Signatures of quantum interference effects on charge transport through a single benzene ring. Angew. Chem. Int. Ed. 52, 3152–3155 (2013).
- Sautet, P. & Joachim, C. Electronic interference produced by a benzene embedded in a polyacetylene chain. Chem. Phys. Lett. 153, 511–516 (1988).
- Su, T. A., Li, H., Steigerwald, M. L., Venkataraman, L. & Nuckolls, C. Stereoelectronic switching in single-molecule junctions. *Nat. Chem.* 7, 215–220 (2015).
- Su, T. Á., Neupane, M., Steigerwald, M. L., Venkataraman, L. & Nuckolls, C. Chemical principles of single-molecule electronics. *Nat. Rev. Mater.* 1, 16002 (2016).
- Ha, Y.-G., Everaerts, K., Hersam, M. C. & Marks, T. J. Hybrid gate dielectric materials for unconventional electronic circuitry. Acc. Chem. Res. 47, 1019–1028 (2014).
- Robertson, J. & Wallace, R. M. High-K materials and metal gates for CMOS applications. *Mater. Sci. Eng. Rep.* 88, 1–41 (2015).
- Simmons, J. G. Generalized formula for the electric tunnel effect between similar electrodes separated by a thin insulating film. J. Appl. Phys. 34, 1793–1803 (1963).
- 12. Trouwborst, M. L. et al. Transition voltage spectroscopy and the nature of vacuum tunneling. *Nano Lett.* **11**, 614–617 (2011).
- Andrews, D. Q., Solomon, G. C., Van Duyne, R. P. & Ratner, M. A. Single molecule electronics: increasing dynamic range and switching speed using crossconjugated species. J. Am. Chem. Soc. 130, 17309–17319 (2008).



- Li, H. et al. Extreme conductance suppression in molecular siloxanes. J. Am. Chem. Soc. 139, 10212–10215 (2017).
- Borges, A., Fung, E. D., Ng, F., Venkataraman, L. & Solomon, G. C. Probing the conductance of the σ-system of bipyridine using destructive interference. J. Phys. Chem. Lett. 7, 4825–4829 (2016).
- Adak, O. et al. Flicker noise as a probe of electronic interaction at metal–single molecule interfaces. Nano Lett. 15, 4143–4149 (2015).
- Tsuji, H., Michl, J. & Tamao, K. Recent experimental and theoretical aspects of the conformational dependence of UV absorption of short chain peralkylated oligosilanes. J. Organomet. Chem. 685, 9–14 (2003).
- Tsuji, H., Terada, M., Toshimitsu, A. & Tamao, K. σσ* transition in anti,cisoid alternating oligosilanes: clear-cut evidence for suppression of conjugation effect by a cisoid turn. J. Am. Chem. Soc. 125, 7486–7487 (2003).
- Bande, A. & Michl, J. Conformational dependence of σ-electron delocalization in linear chains: permethylated oligosilanes. *Chem. Eur. J.* 15, 8504–8517 (2009).
- Solomon, G. C., Herrmann, C., Hansen, T., Mujica, V. & Ratner, M. A. Exploring local currents in molecular junctions. *Nat. Chem.* 2, 223–228 (2010).
- George, C. B., Ratner, M. A. & Lambert, J. B. Strong conductance variation in conformationally constrained oligosilane tunnel junctions. *J. Phys. Chem. A* 113, 3876–3880 (2009).
- Löfås, H., Emanuelsson, R., Ahuja, R., Grigoriev, A. & Ottosson, H. Conductance through carbosilane cage compounds: a computational investigation. J. Phys. Chem. C 117, 21692–21699 (2013).
- Fischer, R., Konopa, T., Ully, S., Baumgartner, J. & Marschner, C. Route Si₆ revisited. J. Organomet. Chem. 685, 79–92 (2003).
- Xu, B. & Tao, N. J. Measurement of single-molecule resistance by repeated formation of molecular junctions. *Science* 301, 1221 (2003).
- Venkataraman, L. et al. Śingle-molecule circuits with well-defined molecular conductance. Nano Lett. 6, 458–462 (2006).
- Reddy, P., Jang, S.-Y., Segalman, R. A. & Majumdar, A. Thermoelectricity in molecular junctions. Science 315, 1568 (2007).
- Paulsson, M. & Datta, S. Thermoelectric effect in molecular electronics. Phys. Rev. B 67, 241403 (2003).
- Bergfield, J. P., Solis, M. A. & Stafford, C. A. Giant thermoelectric effect from transmission supernodes. ACS Nano 4, 5314–5320 (2010).
- Widawsky, J. R., Darancet, P., Neaton, J. B. & Venkataraman, L. Simultaneous determination of conductance and thermopower of single molecule junctions. *Nano Lett.* 12, 354–358 (2012).
- Rincón-García, L., Evangeli, C., Rubio-Bollinger, G. & Agrait, N. Thermopower measurements in molecular junctions. *Chem. Soc. Rev.* 45, 4285–4306
- Tamblyn, I., Darancet, P., Quek, S. Y., Bonev, S. A. & Neaton, J. B. Electronic energy level alignment at metal–molecule interfaces with a GW approach. *Phys. Rev. B* 84, 201402 (2011).

- 32. Yoshizawa, K., Tada, T. & Staykov, A. Orbital views of the electron transport in molecular devices. *J. Am. Chem. Soc.* **130**, 9406–9413 (2008).
- Aradhya, S. V. et al. Dissecting contact mechanics from quantum interference in single-molecule junctions of stilbene derivatives. *Nano Lett.* 12, 1643–1647 (2012)

Acknowledgements G.C.S. and M.H.G. received funding from the Danish Council for Independent Research | Natural Sciences and the Carlsberg Foundation. We thank the National Science Foundation (NSF) for the support of experimental studies under grant no. CHE-1404922 (Ha.L.) and Columbia University's Research Initiatives in Science and Engineering. Y.C., Z.S., T.L. and S.X. are sponsored by the National Natural Science Foundation of China (grant nos 21473113 and 51502173), the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (no. 2013-57), the "Shuguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission (14SG40), the Program of Shanghai Academic/Technology Research Leader (no. 16XD1402700), the National Natural Science Foundation of Shanghai (no. 15ZR1431100), the Ministry of Education of China (PCSIRT_16R49) and the International Joint Laboratory of Resource Chemistry (IJLRC). T.A.S. was supported by an NSF Graduate Research Fellowship under grant no. 11-44155. We thank B. Fowler for mass spectrometry characterization. Single-crystal X-ray diffraction was performed at the Shared Materials Characterization Laboratory (SMCL) at Columbia University. Use of the SMCL was made possible by funding from Columbia University.

Author contributions M.H.G., Ha.L., T.A.S., C.N., L.V. and G.C.S. conceived the idea for the paper. M.H.G. conducted the theoretical calculations under the supervision of G.C.S. Ha.L. did the conductance, noise and thermopower measurements under the supervision of L.V. Y.C., T.A.S., Z.S, D.W.P. and T.L. synthesized and characterized the molecules under the supervision of F.N., He.L., C.N. and S.X. M.H.G., Ha.L., L.V. and G.C.S. analysed the data and wrote the paper with contributions from all authors.

Competing interests The authors declare no competing financial interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0197-9.

Reprints and permissions information is available at http://www.nature.com/reprints

Correspondence and requests for materials should be addressed to G.C.S., L.V.,

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Theory. We model the single-molecule junction conductance as coherent tunnelling using the Landauer-Büttiker scattering formalism. The Landauer transmission is calculated using the equilibrium limit of the non-equilibrium Green's functions (NEGF) approach as described in detail elsewhere $^{34}\!.$ We optimize all molecules in vacuum to a force threshold of 0.01 eV Å⁻¹ for all atoms using DFT with the Perdew-Burke-Ernzerhof correlation-exchange (PBE XC) functional³⁵ and double- ζ plus polarization (DZP) basis set as implemented in the Atomic Simulation Environment (ASE) and GPAW software packages^{36–38}. We use a systematic algorithm for building single-molecule junctions in ASE. We place the tip Au-atoms on the optimized molecules in positions corresponding to a fully extended junction. Based on test optimizations, we use the following parameters: Au-S distance 2.45 Å, Au-S-CH₂ bond angle 110°, Au-S-CH₂-Si dihedral angle $\pm 170^{\circ}$. Based on the position of the tip Au-atoms, we build two four-atom Au pyramids (tetrahedrons) and place these on a $4 \times 4 \times 4$ Au face-centred cubic (fcc) (111) slab to form a single-molecule junction with periodic boundary conditions in all directions. We relax the junction structures to $0.05\,\text{eV}$ Å $^{-1}$ using DFT with the PBE XC functional, DZP basis set for the molecule and double- ζ (DZ) basis set for the Au atoms, two **k**-points in the irreducible part of the Brillouin zone for the x- and y-directions, and one k-point for the z-direction (the transport direction of the junction). The optimized junction structures (the two Au-pyramids and the molecule) are placed between semi-infinite 6×6 Au fcc111 surfaces, and the Landauer transmission is calculated using the NEGF approach as implemented in Atomistix Tool Kit (ATK) software package $^{34,39-41}$. The transmission is calculated using DFT with the PBE XC functional, DZP basis set for the molecule and DZ basis set for the Au atoms, one k-point in the irreducible part of the Brillouin zone for the x- and y-directions, and 200 k-points for the z-direction (the transport direction of the junction). We resolve the interatomic transmission pathways at the Fermi energy as described in detail elsewhere²⁰. The transmission contributions between atoms are plotted as arrows on top of the optimized junction structures in Fig. 2. At any surface across the junction, the sum of arrows reproduces the full transmission. The cross-sectional area of the arrows scales proportionally with the magnitude of the interatomic transmission. We apply a threshold of 5% of the total transmission for the pathways. Pathways between 5% and 20% of the total transmission are plotted in dimmed colours. Note that when there is destructive quantum interference, the magnitude of one pathway can be larger than the total transmission. We use two Au ghost atoms to provide extra basis functions when calculating the transmission through vacuum for Si222-cut as presented in Fig. 2f. Tests show that the extra ghost atom basis functions are only needed for Si-Si gaps over 4.5 Å. Therefore, ghost atoms are not needed for the transmission plotted in Fig. 2b, and consequently we can calculate the transmission pathways as shown in Fig. 2e without ghost atoms. We have tested that the through-space H-Si-Si-H dihedral between the two silyl groups of Si222-cut has no noticeable effect on the transmission.

Synthesis. All reactions were performed in oven-dried or flame-dried round-bottom flasks, unless otherwise noted. The flasks were fitted with rubber septa, and reactions were conducted under a positive pressure of nitrogen or argon, unless otherwise noted. THF, hexane and toluene were obtained from a Schlenk manifold with purification columns packed with activated alumina and supported copper catalyst (Glass Contour). Commercial reagents were used without further purification. Dichlorodimethylsilane and dichlorotetramethyldisilane were purchased from TCL. All other reagents were purchased from Sigma-Aldrich. 1,4-bis(trimethylsilyl)dodecamethylbicyclo[2.2.2]octasilane was synthesized according to previously reported methods²³.

1-chloro-2-(methylthiomethyl)tetramethyldisilane 1: A 500-ml three-neck round-bottom flask was equipped with a stir bar, addition funnel and condenser. Magnesium turnings (14.5 g, 0.6 mol, 2.07 equiv.) were activated with a small crystal of iodine in 10 ml of THF. We subsequently added 190 ml of THF. We added chloromethyl methyl sulfide (0.29 mol, 24.3 ml, 1.00 equiv.) dropwise through the addition funnel over a period of 1 h. During this period, the temperature of the reaction system was kept at 10-20 °C by cooling with an ice-water bath; maintenance of the reaction temperature between 10 and 20 °C is important to the successful generation of the Grignard reagent. After stirring at room temperature for an additional hour, we cannula transferred the solution and passed it through a Schlenk filter to give methylthiomethylmagnesium chloride as an assumed 1.45 M solution in THF, which was subsequently added dropwise to dichlorotetramethyldisilane (0.29 mol, 54.29 g, 1.00 equiv.) in THF (480 ml) at room temperature. The mixture was heated under reflux for 6 h. THF was removed under reduced pressure and hexane was added to the residue. The hexane solution was filtered over Celite through a Schlenk filter into a Schlenk flask. The solvent was removed in vacuo, and the crude compound was vacuum distilled to yield 1 as a light yellow oil (46.30 g, 75%). 1 H NMR (400 MHz, $C_{6}D_{6}$) δ 1.84, 1.69, 0.45, 0.19. ¹³C NMR (101 MHz, C_6D_6) δ 20.39, 19.27, 2.75, -4.50. ²⁹Si NMR (79 MHz, C_6D_6) δ 22.76, -16.50. High-resolution mass spectrometry (HRMS) could not be obtained because of the sensitivity of the compound to water.

Si222: We added 1,4-bis(trimethylsilyl)dodecamethylbicyclo[2.2.2]octasilane²³ (0.557 g, 1.01 mmol, 1 equiv.), tert-BuOK (0.226 g, 2.02 mmol, 2 equiv.) and 18-crown-6 (0.534 g, 2.02 mmol, 2 equiv.) to a 100-ml Schlenk flask, followed by toluene (20 ml). The mixture was stirred overnight to generate the octasilyl dianion. This solution was then added dropwise to a solution of chloromethyl methyl sulfide (0.195 g, 2.02 mmol, 2 equiv.) in 40 ml of toluene cooled to -78 °C in a bath of dry ice plus acetone. The reaction mixture was allowed to warm up to room temperature and stirred for 4 h. The reaction mixture was quenched with 2 M H₂SO₄ and extracted three times with diethyl ether. The organic layers were combined, dried over magnesium sulfate and concentrated in vacuo. The crude residue was purified by silica gel chromatography (gradient from hexanes to 7:3 hexanes:dichloromethane) to give a white solid (213 mg, 40%). H NMR (400 MHz, CDCl₃) δ 2.18 (s, 4H), 2.18 (s, 6H), 0.31 (s, 36H). ¹³C NMR (101 MHz, CDCl₃) δ 22.06, 15.00, -2.65. ^{29}Si NMR (79 MHz, CDCl $_{3}$) δ -39.30, -75.79; HRMS (fast atom bombardment, FAB+): calculated for C₁₆H₄₆S₂Si₈ 526.12, found [M+H]⁺ 527.13. Single crystals for X-ray diffraction were grown from vapour diffusion of methanol into a concentrated solution of Si222 in toluene. The mass spectroscopic data were obtained at the Columbia University mass spectrometry facility using a JEOL JMSHX110A/110A tandem mass spectrometer.

Si2-Si222-Si2: 1,4-bis(trimethylsilyl)dodecamethylbicyclo[2.2.2]octasilane²³ (0.557 g, 1.01 mmol, 1 equiv.), tert-BuOK (0.226 g, 2.02 mmol, 2 equiv.) and 18-crown-6 (0.534 g, 2.02 mmol, 2 equiv.) were added to a 100 ml Schlenk flask, followed by toluene (20 ml) and the mixture was stirred overnight to generate the octasilyl dianion. This solution was then added dropwise to a solution of 1 (0.430 g, 2.02 mmol, 2 equiv.) in 40 ml of toluene cooled to -78 °C in a dry ice/acetone bath. The reaction mixture was allowed to warm up to room temperature and stirred for 4h. The reaction mixture was quenched with 2 M H₂SO₄ and extracted three times with diethyl ether. The organic layers were combined, dried over magnesium sulfate and concentrated in vacuo. The crude residue was purified by silica gel chromatography (gradient from hexanes to 7:3 hexanes:dichloromethane) to give a white solid (360 mg, 47%). ¹H NMR (400 MHz, CDCl₃) δ 2.16 (s, 6H), 1.93 (s, 4H), 0.33 (s, 12H), 0.31 (s, 36H), 0.22 (s, 12H). ¹³C NMR (101 MHz, CDCl₃) δ 20.76, 20.59, -0.54, -0.68, -3.22. ²⁹Si NMR (79 MHz, CDCl₃) $\delta -14.47, -35.06,$ -37.18, -126.44; HRMS (FAB+): calculated for $C_{24}H_{70}S_2Si_{12}$ 758.22, found [M+H]⁺ 759.22. Single crystals for X-ray diffraction were grown from vapour diffusion of methanol into a concentrated solution of Si2-Si222-Si2 in toluene. **Conductance measurements.** We measure the single-molecule conductance using the scanning tunnelling microscope break-junction technique with a custom-built set-up described previously²⁵. Briefly, we drive a gold tip in and out of contact with a gold-on-mica substrate and record the conductance (current/voltage) of the junction as the tip is withdrawn. Upon rupture of the Au contact, a molecule may bridge the gap as evidenced by an additional plateau in the trace of conductance versus displacement. We collect 10,000 to 30,000 such traces, which contain 2,000 data points per nanometre of extension (40 kHz sampling rate), and construct the 1D and 2D conductance histograms without data selection. The histograms are normalized by the number of traces used to construct them. All silanes studied here were introduced into the set-up in a 1,2,4-trichlorobenzene solution with 0.1-1 mM concentration.

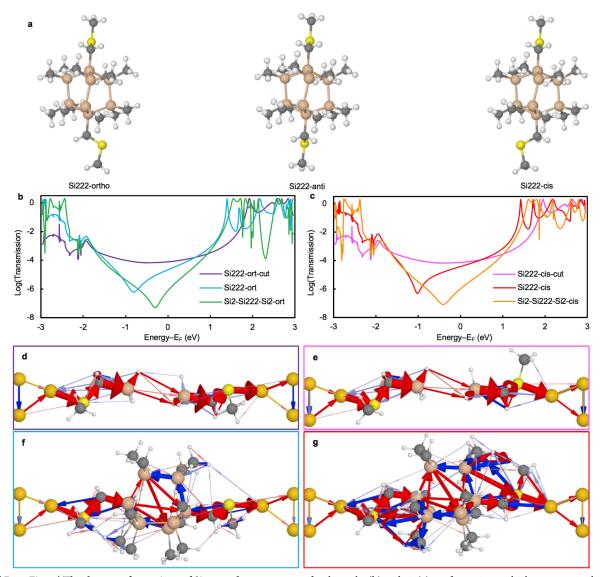
Thermopower measurements. We determined the thermopower of each molecule by performing break-junction measurements with an applied temperature gradient and no applied voltage, as described previously²⁹. Briefly, we heat the substrate and measure the temperature of the substrate and the tip to determine the temperature difference (ΔT) between them. For the work reported here, we have $\Delta T = 0$ K, 27 K and 37 K. We then form a contact between the tip and substrate (at a constant applied voltage) in an environment of the molecules and then break the contact by withdrawing the tip from the substrate at a speed of 20 nm s⁻¹. After a fixed elongation of 2.5 nm, we hold the junction for 5 ms while the bias (750 mV for Si2-Si222-Si2, 45 mV for Si4 and 220 mV for Si222) is applied to measure the junction conductance. We then drop the applied bias voltage to zero and hold the junction for an additional 50 ms while measuring the current. The average current during this time is the average thermoelectric current (*I*). Following this, we turn on the bias again, hold the junction for an additional 5 ms and then pull the junction apart. We determine the conductance (G) just before and after measuring the thermoelectric current. If both conductances are within the molecular histogram peak, as determined from standard conductance measurement, we select the trace. This is done through an automated algorithm. We typically find that 10% of all measured traces (about 5,000-10,000 traces were taken per molecule) have a molecule while the thermoelectric current is measured and are therefore selected. We determine the thermopower as $S = I/(G\Delta T)$ for each junction measured and compile data from hundreds of junctions into the histogram. The Au thermopower of $2\mu V K^{-1}$ is subtracted from the reported data.

Flicker noise measurements. We characterized the conductance noise of the molecular junctions to differentiate between through-bond and through-space charge transport using a method described previously¹⁶. To measure the noise, we pause the elongation procedure for 100 ms (as detailed above for the thermoelectric current measurement) and record the conductance at a bias of 220 mV or 750 mV (for Si222 and Si2-Si222-Si2, respectively) with a 100-kHz sampling rate. We select traces that sustain a molecular junction (as detailed above) during the hold period and calculate the discrete Fourier transform of this data. Two quantities are calculated from the measured conductance while the junction is held for each of these traces: the average conductance (G) and the normalized noise power (power spectrum density (PSD)/*G*). The PSD is obtained from the square of the integral of the discrete Fourier transform of the measured conductance between 100 Hz to 1000 Hz. The lower frequency limit is constrained by the mechanical stability of the set-up. The upper limit is determined by the input noise of the current amplifier. Using these quantities, we create 2D histograms of the normalized noise power against the average conductance. The relation between noise power and conductance is extracted by determining the exponent *n* for which PSD/*G* and *G* are not correlated. We have previously shown that the relationship between flicker noise power and junction conductance follows a power law dependence (PSD proportional to G^n) with the scaling exponent (n) being indicative of the electronic coupling type: n = 1 is characteristic of through-bond coupling whereas n = 2 is characteristic for through-space coupled junctions. Details of these derivations are presented elsewhere 16

Data availability. The data that support the findings of this study are available from the corresponding authors upon reasonable request. Crystallographic data for the structures reported in this paper have been deposited at the Cambridge Crystallographic Data Centre (CCDC) under the deposition numbers CCDC 1571457 (Si222) and CCDC 1571458 (Si2-Si222-Si2). Copies of the data can be obtained free of charge from www.ccdc.cam.ac.uk/data_request/cif.

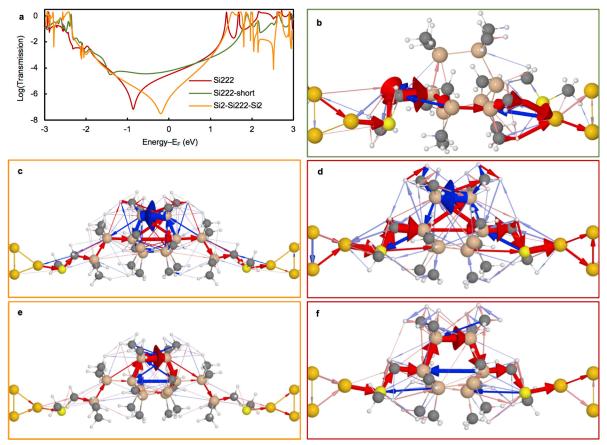
Code availability. The data that support the experimental findings were acquired using a custom instrument controlled by custom software (Igor Pro, Wavemetrics). The software is available from the corresponding author upon reasonable request.

- Brandbyge, M., Mozos, J.-L., Ordejón, P., Taylor, J. & Stokbro, K. Densityfunctional method for nonequilibrium electron transport. *Phys. Rev. B* 65, 165401 (2002).
- 35. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
- Bahn, S. R. & Jacobsen, K. W. An object-oriented scripting interface to a legacy electronic structure code. *Comput. Sci. Eng.* 4, 56–66 (2002).
- Mortensen, J. J., Hansen, L. B. & Jacobsen, K. W. Real-space grid implementation of the projector augmented wave method. *Phys. Rev. B* 71, 035109 (2005).
- Larsen, A. H., Vanin, M., Mortensen, J. J., Thygesen, K. S. & Jacobsen, K. W. Localized atomic basis set in the projector augmented wave method. *Phys. Rev.* B 80, 195112 (2009).
- Soler, J. M. et al. The SIESTA method for ab initio order-N materials simulation. J. Phys. Condens. Matter 14, 2745 (2002).
- Atomistix ToolKit version 2016.3 (QuantumWise A/S, 2016); www.quantum wise.com.
- Virtual NanoLab version 2016.3 (QuantumWise A/S, 2016); www.quantumwise. com
- Markussen, T., Jin, C. & Thygesen, K. S. Quantitatively accurate calculations of conductance and thermopower of molecular junctions. *Phys. Status Solidi B* 250, 2394–2402 (2013).
- Venkataraman, L., Klare, J. E., Nuckolls, C., Hybertsen, M. S. & Steigerwald, M. L. Dependence of single-molecule junction conductance on molecular conformation. *Nature* 442, 904–907 (2006).
- 44. Su, T. A., Li, H., Steigerwald, M. L., Venkataraman, L. & Nuckolls, C. Stereoelectronic switching in single-molecule junctions. *Nat. Chem.* **7**, 215–220 (2015).
- Park, Y. S. et al. Contact chemistry and single-molecule conductance: a comparison of phosphines, methyl sulfides, and amines. J. Am. Chem. Soc. 129, 15768–15769 (2007).
- Kim, T., Vázquez, H., Hybertsen, M. S. & Venkataraman, L. Conductance of molecular junctions formed with silver electrodes. *Nano Lett.* 13, 3358–3364 (2013).
- 47. Meisner, J. S. et al. Importance of direct metal $-\pi$ coupling in electronic transport through conjugated single-molecule junctions. *J. Am. Chem. Soc.* **134**, 20440–20445 (2012).
- Li, H. et al. Silver makes better electrical contacts to thiol-terminated silanes than gold. Angew. Chem. Int. Ed. 56, 14145–14148 (2017).



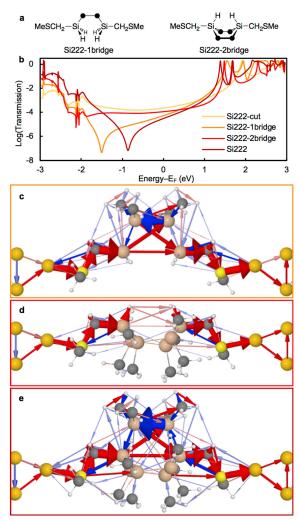
Extended Data Fig. 1 | The three conformations of Si222 and Si2-Si222-Si2, and their transmission data. a, Optimized structures of the three conformations of Si222. b, c, Calculated Landauer transmission

for the ortho (b) and cis (c) conformations, which are very similar to that of the anti-conformation shown in Fig. 2. d-g, Transmission pathways for: (d) Si222-ortho-cut, (e) Si222-cis-cut, (f) Si222-ortho and (g) Si222-cis.

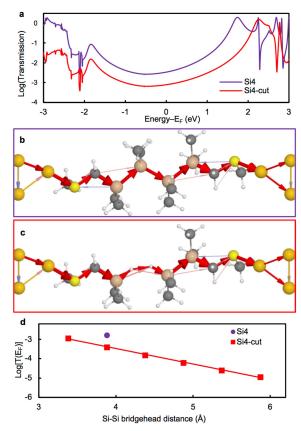


Extended Data Fig. 2 | Transmission data for Si222, Si222-short (a compressed junction of Si222) and Si2-Si222-Si2 anti-conformations. a, Transmission plot; b–d, transmission pathways of (b) Si222-short, (c) Si2-Si222-Si2 and (d) Si222 calculated at $0 \, \text{eV}$ (the Fermi energy); e, Si2-Si222-Si2 calculated at $-0.6 \, \text{eV}$; f, Si222 calculated at $-1.6 \, \text{eV}$. Through-space injection dominates on one side of the molecule in Si222-short at $0 \, \text{eV}$ and a flattened transmission function around

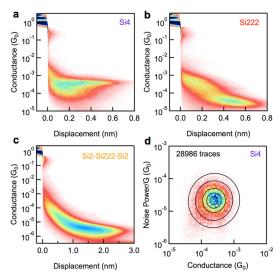
the Fermi energy is observed. This coexistence implies that throughspace injection can change the slope of the transmission (and thus the thermopower), without a substantial change in the magnitude of the conductance. Si2-Si222-Si2 at $-0.6\,\mathrm{eV}$ and Si222 at $-1.6\,\mathrm{eV}$ both show reversed ring current direction in comparison with that calculated at $0\,\mathrm{eV}$, a clear signature of destructive quantum interference.



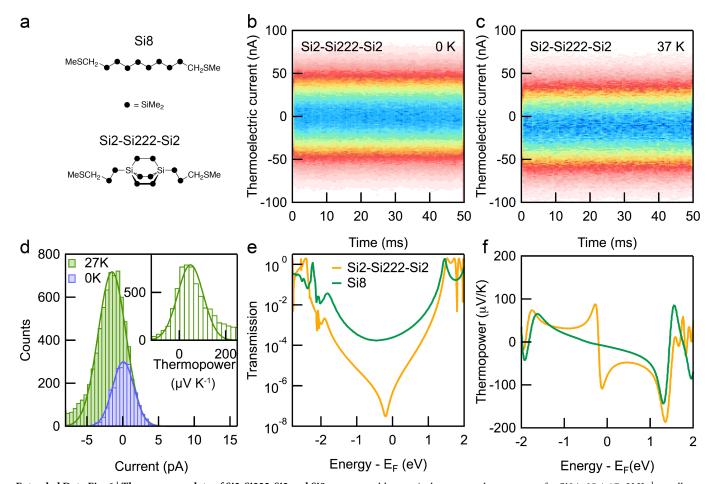
Extended Data Fig. 3 | Transmission of the partially cut versions of Si222. a, Chemical structures of Si222-1bridge and Si222-2bridge. b, Transmission data of Si222 in anti-conformation with different number of bridges being cut. c-e, Transmission pathways are shown for (c) Si222-1bridge (one bridge remains, two were cut), (d) Si222-2bridge (two bridges remain, one was cut) and (e) Si222. The transmission of the Si222-2bridge junction is almost as high as Si222-cut junction where all three bridges are cut, because the bridge where the interference signature appears is cut (d). If we instead cut the other two bridges simultaneously (c, Si222-1bridge), the ring current pathways and the antiresonance in the transmission persist.



Extended Data Fig. 4 | Transmission of linear tetrasilane (Si4) with one bridging unit cut. a, Transmission plot of Si4 and Si4-cut, where one $Si(CH_3)_2$ unit has been cut away and the bridgehead silicon atoms passivated with hydrogen atoms. b, c, Transmission pathways are shown for (b) Si4 and (c) Si4-cut. d, Transmission at the Fermi energy plotted against bridgehead silicon distance of Si4 and Si4-cut. Solid red line is a linear fit to the data of Si4-cut.

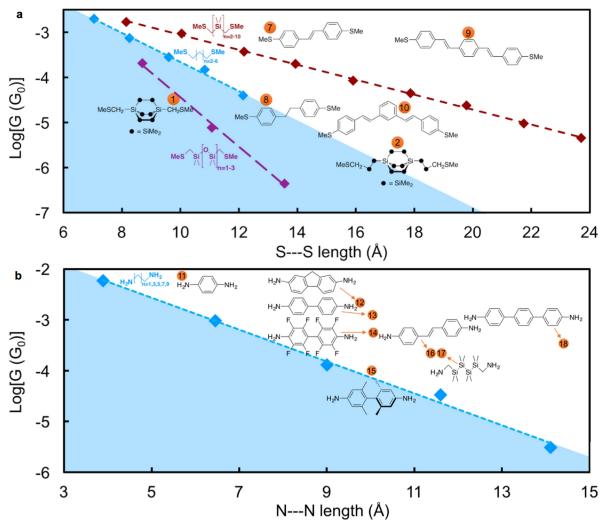


Extended Data Fig. 5 | Experimental 2D conductance versus displacement histograms. a, Si4; b, Si222; c, Si2-Si222-Si2. d, Two-dimensional histogram of normalized flicker noise power against average junction conductance for Si4 along with a 2D Gaussian fit of the data. We see almost no correlation between flicker noise power and the conductance and the noise power scales as $G^{1.1}$.



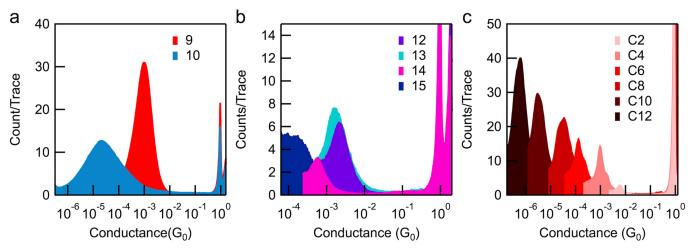
Extended Data Fig. 6 | Thermopower data of Si2-Si222-Si2 and Si8. a, Chemical structures of Si8 and Si2-Si222-Si2. b, c, Two-dimensional histograms of thermoelectric current measured while Si2-Si222-Si2 junctions are held at $\Delta T = 0~\rm K$ and 37 K, respectively. d, Histogram of the measured thermoelectric current for Si8, which has the same number of Si atoms across the molecule as Si2-Si222-Si2, with $\Delta T = 0~\rm K$ and 27 K. Inset: Histogram of thermopower determined from the thermoelectric current for Si8 junctions. After subtracting the thermopower of Au of $2~\rm \mu V~\rm K^{-1}$ (to account for the thermoelectric current between the hot substrate and the

cold set-up), the average thermopower for Si8 is $35\pm17\,\mu\mathrm{V}\ K^{-1}$, smaller than that of Si2-Si222-Si2. e, Transmission curves for Si8 junction along with Si2-Si222-Si2 junction showing different slopes at the Fermi level. f, Thermopower calculated as the slope of the transmission as a function of energy. Theory underestimates the thermopower of both Si2-Si222-Si2 and Si8 by approximately an order of magnitude. Furthermore, the energy alignment between the antiresonance and the Fermi energy is not exact because of inherent errors of DFT 31,42 , which results in the opposite sign of the thermopower at the Fermi energy compared to the experimental value.



Extended Data Fig. 7 | Further comparison of experimental conductance of thiomethyl- and amine-terminated molecules.
a, b, Experimental conductance is plotted against calculated (a) sulfursulfur distance for thiomethyl-linked molecules and (b) nitrogen-nitrogen

distance for a mine linked molecules; dashed lines are linear fits to the data. All conductance values are determined from log-binned conductance histograms created from data taken from references and reproduced in Extended Data Fig. $8^{14,33,43-48}$.



Extended Data Fig. 8 | Logarithmically binned 1D conductance histograms for control molecules. a, Molecules 9 and 10; b, molecules 12 to 15; c, amine-terminated alkanes C2 to C12.



Ballistic molecular transport through twodimensional channels

A. Keerthi^{1,2}, A. K. Geim^{1,2}*, A. Janardanan¹, A. P. Rooney³, A. Esfandiar^{2,4}, S. Hu², S. A. Dar^{1,2}, I. V. Grigorieva¹, S. J. Haigh³, F. C. Wang^{1,5} & B. Radha^{1,2}*

Gas permeation through nanoscale pores is ubiquitous in nature and has an important role in many technologies 1,2. Because the pore size is typically smaller than the mean free path of gas molecules, the flow of the gas molecules is conventionally described by Knudsen theory, which assumes diffuse reflection (random-angle scattering) at confining walls³⁻⁷. This assumption holds surprisingly well in experiments, with only a few cases of partially specular (mirrorlike) reflection known^{5,8-11}. Here we report gas transport through ångström-scale channels with atomically flat walls 12,13 and show that surface scattering can be either diffuse or specular, depending on the fine details of the atomic landscape of the surface, and that quantum effects contribute to the specularity at room temperature. The channels, made from graphene or boron nitride, allow helium gas flow that is orders of magnitude faster than expected from theory. This is explained by specular surface scattering, which leads to ballistic transport and frictionless gas flow. Similar channels, but with molybdenum disulfide walls, exhibit much slower permeation that remains well described by Knudsen diffusion. We attribute the difference to the larger atomic corrugations at molybdenum disulfide surfaces, which are similar in height to the size of the atoms being transported and their de Broglie wavelength. The importance of this matter-wave contribution is corroborated by the observation of a reversed isotope effect, whereby the mass flow of hydrogen is notably higher than that of deuterium, in contrast to the relation expected for classical flows. Our results provide insights into the atomistic details of molecular permeation, which previously could be accessed only in simulations $\hat{1}^{10,14}$, and demonstrate the possibility of studying gas transport under controlled confinement comparable in size to the quantum-mechanical size of atoms.

Knudsen theory provides a comprehensive description of gas flow in the regime in which molecules collide mostly with confining walls rather than each other. Despite its universal adoption, it relies on certain assumptions, including fully diffusive surface scattering^{3,5,7}. Several new experimental systems with nanoscale channels have recently been introduced, including carbon nanotubes^{8,11} and nanoporous films made from graphene^{15–19}, graphene oxide^{19,20} and other two-dimensional (2D) materials^{21,22}. Several anomalies in gas-permeation properties have been reported, relative to expectations, which in certain cases^{8,11} were difficult to reconcile within the classical theory. In particular, the observation of fast gas flows through narrow carbon nanotubes was attributed to a combination of specular and diffusive scattering^{8,11}. Unfortunately, the exact dimensions and structure of these new systems are often insufficiently controlled, which makes it difficult to compare the observed behaviour with a large and growing body of molecular dynamics simulations 10,14. The analysis is further complicated by the poorly understood effects of nanotube curvature²³ and, especially, the presence of adsorbates (such as hydrocarbons), which universally cover surfaces that are not under ultrahigh vacuum^{9,24–26}.

In this work, we report gas transport through angström-scale slit-like channels with walls made from cleaved crystals of graphite, hexagonal boron nitride (h-BN) or molybdenum disulfide (MoS₂). These three materials were chosen as archetypal examples of crystals that can be exfoliated down to monolayers and provide atomically flat surfaces that are stable under ambient conditions¹². The nanochannels were fabricated following the procedure described in Methods section 'Making 2D channels. In brief, two thin (roughly 10-100 nm) crystals of the above materials were prepared by exfoliation to serve as bottom and top walls of a channel. A third, thinner crystal was plasma-etched to contain long narrow trenches (Fig. 1). It served as a spacer between the top and bottom walls. The three crystals were assembled on top of each other as shown in Fig. 1a, held together by van der Waals forces¹². The slit height h was determined by the van der Waals thickness of the spacer crystal and could be chosen to be just one atomic layer up to as many as required. In all of the measurements reported here, the trench width w that defines the width of the resulting channels was about 130 nm, chosen to be sufficiently large to allow accurate measurements of a gas flow, but not large enough to allow sagging¹³. To increase the measurement accuracy, we often used many slits in parallel (typically 200), but some experiments were also done using individual channels. Their length L was defined by lithography and ranged from about 1 μ m to more than 15 μ m (Extended Data Fig. 1). An example of our 2D slits is shown in Fig. 1a, b, where MoS₂ was intentionally used as the building material to provide high contrast for transmission electron microscopy (TEM) imaging. This channel may be viewed as if a single atomic plane was removed from a bulk MoS2 crystal, resulting in a pair of edge dislocations and a 2D empty space with $h \approx 6.7$ Å. We made and studied slits using different wall materials (see Methods section 'Crosssectional imaging of 2D channels'), but used mainly graphene spacers, which allowed a slit height $h = N \times a$ in multiples of the graphene van der Waals thickness $a \approx 3.4$ Å, where N is the number of layers.

For measurements of gas transport through the 2D channels, we made devices as shown schematically in Fig. 1c. The van der Waals assembly was placed to seal a micrometre-sized opening in a silicon wafer (Extended Data Figs. 1, 3). The wafer separated two containers, one of which had a gas under variable pressure P whereas the other was a vacuum chamber equipped with a mass spectrometer. Unless otherwise stated, we used helium as a test gas. The applied pressure P was increased slowly to 200 mbar, which corresponds to the mean free path always being larger than 0.7 µm and typical Knudsen numbers of more than 10^4 . Examples of the measured flow rates Q as a function of P are shown in Fig. 1d. As a reference, we made devices using the same fabrication procedures but without etching trenches in the spacer crystal or with partial trenches that did not connect the two containers. Those devices exhibited no discernable helium permeation, confirming that the 2D channels were the only pathway between the two containers. The minimal spacers were monolayer MoS₂ ($h \approx 6.7$ Å; as in Fig. 1) and bilayer graphene (N=2; $h\approx 6.8$ Å).

Helium transport through our 2D channels was found to depend strongly on the wall material. In Fig. 1d we show that devices of exactly the same geometry ($h = 5a \approx 17 \text{ Å}$) exhibited flow rates Q that were two

¹School of Physics and Astronomy, University of Manchester, Manchester, Manchester, UK. ²National Graphene Institute, University of Manchester, Manchester, UK. ³School of Materials, University of Manchester, Manchester, UK. ⁴Department of Physics, Sharif University of Technology, Tehran, Iran. ⁵Chinese Academy of Sciences Key Laboratory of Mechanical Behavior and Design of Materials, Department of Modern Mechanics, University of Science and Technology of China, Hefei, China. *e-mail: geim@manchester.ac.uk; radha.boya@manchester.ac.uk

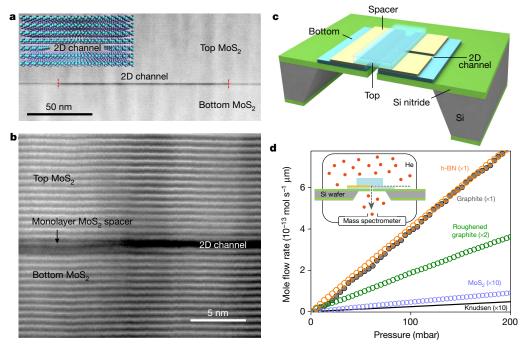


Fig. 1 | Helium gas transport through ångström-scale slits. a, Schematic (inset) and TEM micrograph (main panel) of a 2D channel assembled from MoS_2 crystals. The channel is seen in black in the main panel; for clarity, its edges are marked with red ticks. The monolayer spacer appears darker with respect to the top and bottom crystals because of different in-plane orientations. The contrast ripples running vertically are a result of the curtaining effect that occurs during ion-beam polishing³². Micrographs of slits made from other 2D materials are provided in Extended Data Fig. 2 and ref. ¹³. b, High-magnification image of the channel in a near its left edge. Each bright horizontal line corresponds to monolayer MoS_2 . c, Schematic of our experimental devices.

The tri-crystal assembly (cyan and yellow) covers an aperture in a silicon nitride membrane (green) prepared on top of a silicon wafer (grey). **d**, Comparison of helium permeation through 2D channels of the same height (N=5), but with walls made from different crystals (as indicated by the labels). All of the devices here are single-channel, with $L=1-6~\mu m$. The (mole) flow rates at room temperature (296 \pm 3 K) are normalized per channel length and, for legibility, multiplied by the factors shown. The flow expected for Knudsen diffusion is shown by the solid black line close to the MoS₂ data. Inset, our measurement set-up. The arrow indicates the gas flow direction.

orders of magnitude larger for graphene and h-BN walls than for MoS_2 walls. In our case, where the mean free path of helium atoms is much larger than h, Knudsen theory predicts a mass flow rate of 5,6

$$Q_{\rm K} = \alpha P \left(\frac{m}{2\pi RT}\right)^{1/2} wh \tag{1}$$

where m is the atomic mass of the gas being transported, R is the gas constant, T is the temperature and α is the transmission coefficient. For long and narrow rectangular channels, α can be approximated as (see Methods section 'Entry effects')

$$\alpha \approx \frac{h}{L} \ln \left(\frac{4w}{h} \right) \tag{2}$$

According to these equations, and converting between mass and mole flow rates, the N=5 channels in Fig. 1d should have flow rates Q shown by the solid black line. The Knudsen prediction holds (within a factor of two) only for MoS₂ walls. The other 2D slits have much higher Q. Similar disparity was observed for other values of h (see below). The large aspect ratios ($w/h \approx 100$) imply that gas molecules collide mostly with the top and bottom walls and therefore that scattering at side walls has a relatively small role, in agreement with our experimental observation that channels made using different side-wall materials but similar h exhibited similar Q (see, for example, channels in Fig. 1 and Extended Data Fig. 2b).

To quantify the observed enhancement with respect to Knudsen theory, we introduce an enhancement coefficient $K = Q/Q_K$. This allows us to summarize our findings by plotting results for more than 70 devices (Fig. 2a). Within the data scatter, the gas flow through MoS₂ channels is well described by equations (1) and (2). By contrast, all of the graphite

and h-BN devices exhibit strong enhancement (K>100). There is a clear tendency for smaller K with increasing N, with K decreasing to about 10 and 3 for N=12 and 25, respectively. This trend is not surprising because larger channels are expected eventually to follow the standard behaviour. The enhancement effect is so strong that graphene and h-BN slits with $h\approx 1.4$ nm exhibited a helium flux roughly ten times larger than the 9-nm slits, in contrast to equations (1) and (2), which suggest a difference by a factor of about ten but in the opposite direction.

The flow enhancement that we observe can be attributed to specular reflection of helium atoms off atomically flat walls, which should result in the breakdown of Knudsen's approximation. This possibility was first considered by Smoluchowski⁴, who modified the theory by introducing the fraction f of diffusely reflected molecules. In this case, the transmission coefficient α should—in the first approximation 4,27 —be multiplied by the factor (2-f)/f. Using the Smoluchowski model 4,5,10 , our data yield, for example, $f\approx 0.2$ for the 4-nm graphite and h-BN channels and $f\approx 1$ for all MoS $_2$ devices. The largest K values in Fig. 2a suggest values of f very close to zero, that is, specular reflection. Even without taking Smoluchowski theory into account, our data for graphene channels with N=4 yield practically perfect transmission. This limit $(\alpha=1)$ is plotted in Fig. 2a (dashed curve) and corresponds to the free molecular flow through an aperture of size $w\times h$. The agreement that we observe means that helium atoms pass through our long slit-like channels ballistically, without losing their momenta.

To further substantiate the observed frictionless flow, we carried out two additional experiments. First, we made graphene channels with roughened bottom walls using a short exposure to oxygen plasma (Methods section 'Making 2D channels'). This roughness suppressed

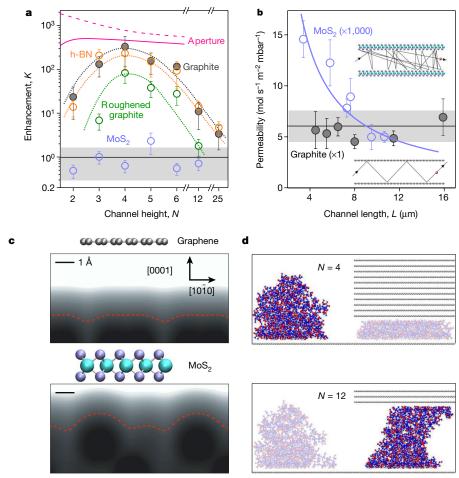


Fig. 2 | Enhanced gas flow through 2D channels. a, Enhancement factor K observed for different wall materials and channel heights (symbols). The channels could also have different L. The dotted curves are guides to the eye. MoS_2 channels exhibit no enhancement within our data scatter. The dashed pink curve indicates the K behaviour expected from equation (1) for ideal transmission. The solid pink curve is also for $\alpha=1$, but takes into account the finite size d of helium atoms (Methods section 'Entry effects', Extended Data Fig. 7). b, Dependence of helium transport on L, with N=4. The permeability is the flow rate divided by $P\times w\times h$. In the case of MoS_2 walls, the permeability follows the 1/L dependence expected for diffusive transport (the purple data are multiplied by a factor of 1,000); the purple curve is a 1/L fit. No length dependence is found for graphene channels, a clear signature of frictionless transport.

The insets illustrate diffusive (top) and specular (bottom) scattering. Error bars in $\bf a$ and $\bf b$ show statistical errors from our measurements, using at least two devices for each N in the case of $\bf a$. The shaded areas in $\bf a$ and $\bf b$ indicate the standard error for the datasets shown by the purple symbols for MoS₂ in $\bf a$ and the black symbols for graphite in $\bf b$. $\bf c$, Intrinsic roughness of atomically flat surfaces. Grey scale, electron density near graphene and MoS₂ surfaces; red curves, depth accessible for helium atoms with thermal energies (for details, see Methods section 'Intrinsic surface roughness of 2D crystals'). $\bf d$, Self-cleansing of 2D slits. Favourable and unfavourable positions (bright and faint, respectively) are illustrated for a polymer molecule inside slits with different N.

the helium flow by an order of magnitude, albeit insufficient to recover the Knudsen description (Figs. 1d, 2a). Non-zero specularity can be attributed to the fact that the oxygen plasma tends to etch holes in graphite so that a large portion of its surface remains flat²⁸. However, the clearest evidence for ballistic transport comes from the second set of experiments, in which we used channels of the same height h=4a (strongest enhancement) but with different L. In Fig. 2b we show that the gas flow through the channels with MoS₂ walls decreases with L and exhibits the $Q \propto 1/L$ dependence prescribed by the standard theory. By contrast, for graphene slits the gas flow is independent of L, a hallmark of ballistic transport.

The considerable difference between the devices with MoS₂ and graphene or h-BN walls is surprising. Cleaved MoS₂ may not be as perfect as graphene but its surface still contains few defects²⁹, especially in comparison with roughened graphene. We believe that the difference arises from a finite roughness of ideal, atomically flat surfaces. In Fig. 2c we show that MoS₂ exhibits relatively strong corrugations, reaching around 1 Å in height; incoming helium atoms should therefore be able to 'see' this roughness because its scale is comparable to both the kinetic diameter ($d \approx 2.6$ Å) and de Broglie wavelength ($\lambda_B \approx 0.5$ Å) of helium

at room temperature. The former is a semi-classical notion, but, none-theless, represents the quantum-mechanical size of the electron cloud around helium nuclei³⁰. Graphene and h-BN surfaces are much flatter on this scale (Fig. 2c, Extended Data Fig. 4), suggesting more specular reflection.

The above experiments cannot distinguish between the effects of d and λ_B . To determine whether quantum effects contribute to the observed specular reflection, we compared hydrogen and deuterium permeation. These isotopes have the same d and the same interaction with the walls but different λ_B . Equation (1) suggests that Q should be a factor of $\sqrt{2}$ larger for deuterium, independently of the experimental details (such as the channel geometry or f). This benchmark dependence was validated using micrometre-sized reference apertures (Methods section 'Gas transport measurements'). By contrast, our graphene channels with N=4 exhibited a flow rate Q about $30\% \pm 10\%$ smaller for deuterium than for hydrogen. This finding demonstrates that matter-wave effects contribute to the specular reflection, leading to its suppression for the heavier atoms (deuterium) because they have a shorter λ_B and see an atomic landscape that is rougher than that seen by hydrogen.

Finally, we discuss the changeover from ballistic to classical (diffusive) transport, which is observed for relatively large $h \approx 4-8$ nm (Fig. 2a). We believe the underlying reason for the transition is that such large channels are no longer atomically flat, owing to hydrocarbon contamination. Cleaved surfaces are rapidly covered with various adsorbates even under cleanroom conditions²⁴⁻²⁶. This is particularly obvious in high-resolution TEM where only tiny areas of graphene devoid of hydrocarbons can be found²⁴. However, when two atomically flat surfaces are brought together, this contamination is known to aggregate into well-separated pockets (so-called contamination bubbles), outside of which the attached surfaces become atomically clean, free of any adsorbates 12,25. This self-cleansing is likely to start to occur at a finite separation between 2D crystals because individual polymer molecules tend to arrange themselves into clumps of a few nanometres in height (Fig. 2d, Methods section 'Self-cleansing of 2D slits'). Confining such clumps between parallel walls reduces their configurational entropy that competes with an energy gain due to adhesion (Extended Data Fig. 5). The latter is a surface effect whereas the former is determined by the volume of the clumps, which implies that the squashed clumps become energetically unfavourable at small h and should be squeezed out. Our molecular dynamics simulations (Methods section 'Self-cleansing of 2D slits') show that polymer molecules prefer to sit outside narrow channels, whereas an interior position is more favourable for larger h (Fig. 2d). With increasing N, no sharp cut-off is expected for the self-cleansing mechanism, but it should become increasingly less efficient, which would explain the observed decrease in K at large N in Fig. 2a. As for the decrease in K observed at the other end (N=2 and 3), we speculate that it could be due to another size effect. The space that is not filled with electron clouds and is therefore available for molecular transport is about 6.7 Å wide, only slightly larger than $d \approx 2.6$ Å (Extended Data Fig. 6). In combination with a de Broglie wavelength of $\lambda_{\rm B} \approx 0.5$ Å, which effectively increases d, this effect should decrease the number of atoms that can enter the narrow slits, especially at shallow incident angles (solid curve in Fig. 2a, Methods section 'Entry effects'). Interplay between the mechanisms that suppress molecular flow at large and small N leads to the observed dome-shaped dependence. Further theory and simulations are needed to analyse the size effects in detail.

To conclude, our work offers new understanding for many previous predictions, calculations and observations. For example, it seems to reconcile widely varying results for molecular transport through carbon nanotubes. A strongly enhanced gas flow^{8,11} was reported for sub-2-nm nanotubes, whereas no such enhancement was found using wider tubes, in conceptual agreement with the transition from ballistic to diffusive transport reported here, due to the increasing role of hydrocarbon contamination in larger channels. We also note an analogy with ballistic electron transport in metallic systems. These systems can exhibit a finite electrical resistance even in the absence of electron scattering, which is known as the point contact resistance³¹. This effect is described by the Sharvin formula³¹, which is equivalent to equation (1) with $\alpha = 1$ but for a charge rather than a gas flow. Ballistic effects in molecular transport have not previously been considered, even in theory; the ångström-scale channels that we have demonstrated now make this regime accessible experimentally.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0203-2.

Received: 19 December 2017; Accepted: 14 May 2018; Published online 20 June 2018.

 Baker, R. W. Membrane Technology and Applications 2nd edn (John Wiley & Sons, Chichester, 2004).

- Lu, G. Q. & Zhao, X. S. Nanoporous Materials: Science and Engineering (Imperial College Press. London. 2004).
- Knudsen, M. Die Gesetze der Molekularströmung und der inneren Reibungsströmung der Gase durch Röhren. Ann. Phys. 333, 75–130 (1909).
- v. Smoluchowski, M. Zur kinetischen Theorie der Transpiration und Diffusion verdünnter Gase. Ann. Phys. 338, 1559–1570 (1910).
- Steckelmacher, W. A review of the molecular flow conductance for systems of tubes and components and the measurement of pumping speed. *Vacuum* 16, 561–584 (1966).
- Livesey, R. G. in Foundations of Vacuum Science and Technology (ed. Lafferty, J. M.) Ch. 2 (John Wiley & Sons, New York, 1998).
- Lei, W., Rigozzi, M. K. & McKenzie, D. R. The physics of confined flow and its application to water leaks, water permeation and water nanoflows: a review. Rep. Prog. Phys. 79, 025901 (2016).
- 8. Holt, J. K. et al. Fast mass transport through sub-2-nanometer carbon nanotubes. *Science* **312**, 1034–1037 (2006).
- Agrawal, A. & Prabhu, S. V. Survey on measurement of tangential momentum accommodation coefficient. J. Vac. Sci. Technol. A 26, 634–645 (2008).
- Bhatia, S. K., Bonilla, M. R. & Nicholson, D. Molecular transport in nanopores: a theoretical perspective. *Phys. Chem. Chem. Phys.* 13, 15350–15383 (2011).
- Majumder, M., Chopra, N. & Hinds, B. J. Mass transport through carbon nanotube membranes in three different regimes: ionic diffusion and gas and liquid flow. ACS Nano 5, 3867–3877 (2011).
- Geim, A. K. & Grigorieva, I. V. Van der Waals heterostructures. Nature 499, 419–425 (2013).
- Radha, B. et al. Molecular transport through capillaries made with atomic-scale precision. Nature 538, 222–225 (2016).
- Zhang, W.-M., Meng, G. & Wei, X. A review on slip models for gas microflows. Microfluid. Nanofluidics 13, 845–882 (2012).
- Koenig, S. P., Wang, L., Pellegrino, J. & Bunch, J. S. Selective molecular sieving through porous graphene. Nat. Nanotechnol. 7, 728–732 (2012).
- Berry, V. Impermeability of graphene and its applications. Carbon 62, 1–10 (2013).
- Čelebi, K. et al. Ultimate permeation across atomically thin porous graphene. Science 344, 289–292 (2014).
- Wang, L. et al. Molecular valves for controlling gas phase transport made from discrete ångström-sized pores in graphene. Nat. Nanotechnol. 10, 785–790 (2015)
- Kim, H. W. et al. Selective gas transport through few-layered graphene and graphene oxide membranes. Science 342, 91–95 (2013).
- Li, H. et al. Ultrathin, molecular-sieving graphene oxide membranes for selective hydrogen separation. Science 342, 95–98 (2013).
- Liu, G., Jin, W. & Xu, N. Two-dimensional-material membranes: a new family of high-performance separation membranes. *Angew. Chem. Int. Ed.* 55, 13384–13397 (2016).
- Wang, L. et al. Fundamental transport mechanisms, fabrication and potential applications of nanoporous atomically thin membranes. *Nat. Nanotechnol.* 12, 509–522 (2017).
- Falk, K. et al. Molecular origin of fast water transport in carbon nanotube membranes: superlubricity versus curvature dependent friction. Nano Lett. 10, 4067–4073 (2010).
- Gass, M. H. et al. Free-standing graphene at atomic resolution. Nat. Nanotechnol. 3, 676–681 (2008).
- Haigh, S. J. et al. Cross-sectional imaging of individual layers and buried interfaces of graphene-based heterostructures and superlattices. *Nat. Mater.* 11, 764–767 (2012).
- Li, Z. et al. Effect of airborne contaminants on the wettability of supported graphene and graphite. Nat. Mater. 12, 925–931 (2013).
- Arya, G., Chang, H.-C. & Maginn, E. J. Knudsen diffusivity of a hard sphere in a rough slit pore. *Phys. Rev. Lett.* 91, 026102 (2003).
- You, H. X., Brown, N. M. D. & Al-Assadi, K. F. Radio-frequency plasma etching of graphite with oxygen: a scanning tunnelling microscope study. Surf. Sci. 284, 263–272 (1993).
- Addou, R., Colombo, L. & Wallace, R. M. Surface defects on natural MoS₂. ACS Appl. Mater. Interfaces 7, 11921–11929 (2015).
- Mehio, N., Dai, S. & Jiang, D. Quantum mechanical basis for kinetic diameters of small gaseous molecules. J. Phys. Chem. A 118, 1150–1154 (2014).
- Sharvin, Y. V. A possible method for studying Fermi surfaces. Sov. Phys. JETP 21, 655–656 (1965).
- Giannuzzi, L. A. & Stevie, F. A. (eds) Introduction to Focused Ion Beams: Instrumentation, Theory, Techniques and Practice (Springer, New York, 2005).

Acknowledgements This work was supported by the European Research Council, Lloyd's Register Foundation, the EU Graphene Flagship and the Royal Society. B.R. acknowledges a Leverhulme Early Career Fellowship, a L'Oréal Fellowship for Women in Science and EPSRC grant EP/R013063/1. F.C.W. acknowledges support from the National Natural Science Foundation of China (11772319 and 11572307) and the Shanghai Supercomputer Center. S.J.H. and A.P.R. were funded by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreements ERC-2016-STG-EvoluTEM-715502 and DISCOVERER-2017 737183), the US Defence Threat Reduction Agency (HDTRA1-12-1-0013) and the EPSRC (EP/P009050/1 and EP/K016946/1).



Reviewer information *Nature* thanks L. Bocquet and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.K.G. and B.R. designed and directed the project. A.K., B.R., A.J., S.H., A.E and S.A.D. fabricated the devices. A.K., B.R. and A.J. performed the measurements and their analysis. F.C.W. provided theoretical support. A.P.R. and S.J.H. carried out TEM imaging. A.K.G., B.R., F.C.W., A.K. and I.V.G. wrote the manuscript. All authors contributed to discussions.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0203-2.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.K.G. and B $\mbox{\it R}$

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Making 2D channels. Our devices were made following the microfabrication procedures illustrated in Extended Data Fig. 1. First, a free-standing silicon nitride (SiN_x) membrane with dimensions of about 100 $\mu m \times 100~\mu m$ was prepared, starting with the standard Si wafer covered with a 500-nm-thick layer of SiN_x. This was done using photolithography and wet etching. A rectangular hole about 3 $\mu m \times 26~\mu m$ in size was then made in this membrane using photolithography and dry etching. Next a thin (roughly 10–30 nm) crystal of graphite, h-BN or MoS_2 was mechanically exfoliated and transferred^{12,13} onto the membrane to cover the opening (Extended Data Fig. 1a). This crystal served as the bottom layer in our trilayer-crystal assembly. Following the transfer, the rectangular hole was extended into the bottom layer using the SiN_x membrane as a mask for dry etching from the back side of the Si wafer (Extended Data Fig. 1b). To this end, oxygen plasma was used for etching graphite, whereas h-BN and MoS_2 were etched in a mixture of CHF_3 and oxygen.

To make the second (spacer) layer, 2D crystals of graphene or MoS_2 were exfoliated onto an oxidized Si wafer (300 nm of SiO_2). Crystals of a chosen thickness were then etched into stripes about 130 nm wide and separated by the same distance. This was done using electron-beam lithography (polymethyl methacrylate (PMMA) with a molecular weight of 950,000 was used as a resist) and plasma etching. The PMMA mask was removed by mild sonication in acetone. The resulting stripes were transferred on top of the bottom layer as shown in Extended Data Fig. 1c, d. Next, a relatively thick (roughly 100 nm) crystal of graphite, h-BN or MoS_2 was dry-transferred 13,33 on top of the two-layer assembly so that it covered the rectangular hole and partially overlapped with the spacer stripes. After each layer transfer we annealed our assembly in 10%-hydrogen-in-argon at 300– $400\,^{\circ}$ C for 3 h. The annealing step was critical for the cleanliness of the final devices, to avoid the channels becoming blocked with PMMA residue.

For the experiment that used roughened channels, the bottom graphite crystal was exposed briefly to oxygen plasma to remove approximately three layers of graphene. This was done before transferring the spacer layer. To define the length L of the final channels, a metal mask (5 nm Cr/50 nm Au) was placed by photolithography on top of the final assembly as shown in Extended Data Figs 1e, f. Dry etching through this mask not only allowed us to control L accurately, but also opened the entries of the channels if they were blocked accidentally by the thin (less than 10 nm) edges of the top crystal, which tended to sag inside the channels 13 . For single-channel devices used in some experiments (see, for example, Fig. 1d), extra nanocavities were created around the main channel (Extended Data Fig. 1d) to prevent the formation of bubbles that collect hydrocarbons and other contaminants 12,25,33 and so could block individual channels. These cavities were arranged perpendicular to the main channel (Extended Data Fig. 1d) and did not contribute to gas transport through the final devices.

Cross-sectional imaging of 2D channels. For scanning transmission electron microscopy (STEM) and high-angle annular dark-field (HAADF) imaging, we made thin cross-sectional lamellae by implementing an in situ lift-out procedure 25,34 . Lamellae were cut out perpendicular to the capillary axis by high-precision site-specific milling in Helios Nanolab DualBeam 660, which incorporates scanning electron microscope and focused ion-beam columns. Platinum was deposited using the ion beam to weld the lamella to a micromanipulator, which was then lifted from the substrate. Once transferred to a specialist OmniProbe TEM grid, the lamella foil was thinned to less than 100 nm and then polished further to electron transparency, using 5-kV and subsequently 2-kV ion milling. High-resolution STEM and HAADF images were acquired in an aberration-corrected microscope (FEI Titan G2 80-200 kV) using a probe convergence angle of 21 mrad, a HAADF inner angle of 48 mrad and a probe current of about 70 pA. To ensure that the electron beam was parallel to the 2D channels, it was aligned using the relevant Kikuchi bands of the silicon substrate and the assembled 2D crystals.

Gas transport measurements. Gas permeation through our 2D slits was studied in the steady-state flow regime using the set-up shown in Extended Data Fig. 3a. The SiN_x/Si wafer containing a capillary device was sealed using O-rings to separate two oil-free vacuum chambers. Standard vacuum components were used to make the two-chamber assembly sketched in Extended Data Fig. 3a. The chambers were evacuated before every experimental run. One of the chambers was equipped with an electrically controlled dosing valve that provided the pressure *P* inside, which was monitored by a pressure gauge. The applied pressure was varied slowly to avoid any time delays or hysteresis. The top (entry) side of our devices was facing this chamber (Extended Data Fig. 3a). The devices were sufficiently robust to with stand P up to 2 bar. The other chamber was maintained at a pressure of around 10^{-6} bar, and connected to a mass spectrometer. For measurements of helium, hydrogen and deuterium flows, we used a calibrated helium-leak detector (INFICON UL200) as the mass spectrometer. The leak detector measures the flow rates in units of mbar l s⁻¹, which are straightforward to convert into units of mol s⁻¹ using the ideal-gas equation. All the measurements were done at room temperature ($T\!=\!296\pm3$ K), as measured by a probe mounted close to the device.

The measurement set-up was checked thoroughly for possible leaks using control devices that were prepared following the same fabrication procedures but without complete 2D channels (see the main text). These control devices exhibited no discernible He leak, which demonstrates that the 2D channels were the only possible permeation path for the gases tested. To assure quantitative accuracy of our measurements, we also prepared reference devices containing round apertures made in SiN_x membranes and tested their conductance with respect to helium. The measured flow values were indistinguishable from those given by the Knudsen equation (Extended Data Fig. 3b). For the other gases, the apertures were used to calibrate the sensitivity of our mass spectrometer and to measure the isotope effect, in which the mass flow of deuterium was observed to be a factor of $\sqrt{2}$ higher than that of hydrogen, as required by equation (1).

Intrinsic surface roughness of 2D crystals. To compare the flatness of different 2D crystals that served as the walls of the slits, we used density functional theory (DFT) to calculate near-surface electron-density profiles. The results (Extended Data Fig. 4) illustrate that graphene and h-BN are atomically flatter than MoS₂, as is generally expected. For further analysis of the surface roughness, we use the criterion of the so-called thermal exclusion surface³⁵. This criterion suggests that He atoms effectively 'feel' the surface at a critical density of about 0.03 electron per Å³. Incident atoms with a kinetic energy equivalent to their thermal energy (298 K) cannot penetrate beyond this isosurface, shown by the red curves in Extended Data Fig. 4. The DFT analysis was carried out using the CP2K program³⁶ and the PBE exchange-correlation functional³⁷. The energy cut-off for plane-wave expansions was set at 600 Ry. Gaussian basis sets for the double-zeta valence-polarized (DZVP) quality³⁸ and Goedecker-Teter-Hutter pseudopotentials³⁹ were used in the calculations. Periodic boundary conditions were applied and the vacuum region was set to have a thickness of 40 Å. The electron-density contours were analysed using Multiwfn $^{40}\!.$

We also carried out molecular dynamics (MD) simulations in an attempt to explain the substantial difference in molecular transport through MoS2 and graphite channels that we observed. To this end, we constructed nanochannels with potential-energy surfaces obtained using the classical force-field parameters⁴¹ These surfaces are qualitatively similar to the DFT surfaces shown in Extended Data Fig. 4. The nanochannels connect two reservoirs, one of which was maintained at a pressure of 200 mbar by adjusting the number of helium atoms inside. The other was kept empty to model gas transport driven by a pressure gradient. The gas flow through graphite channels was notably higher than that through the MoS₂ channels with rougher walls, as expected. However, graphite and MoS2 channels both exhibited a finite f and a clear dependence of the helium flux on L (that is, no ballistic transport could be seen in these MD simulations even for graphite walls), in contrast to the experiment. This indicates limitations of classical MD simulations in describing specular surface scattering or, at least, insufficient knowledge of the details of the interaction of helium with atomically flat surfaces. Moreover, our experiments with hydrogen isotopes show that not only the interaction but also the molecular mass are important for allowing perfect specular reflection, which indicates a contribution from quantum effects, as discussed in the main text. Accordingly, we limited our efforts to explain the results using classical MD

Self-cleansing of 2D slits. Self-cleansing 12,13,25 of interfaces during van der Waals assembly has been studied extensively over the past five years. As argued in the main text, similar self-cleansing processes should take place if two atomically flat surfaces are in direct contact or at finite separation. To model this, we take PMMA on graphene as an archetypal example of a poorly mobile adsorbate that is often found on graphene and other surfaces $^{24-26}.\ For\ a\ PMMA$ molecule confined inside an angström-scale slit, the total free energy has two main contributions: adhesion energy with the two walls and configurational entropy. The former tends to keep PMMA inside whereas the latter increases the total energy if PMMA molecules are flattened and, therefore, pushes them out. For strong confinement (small *h*), the configurational entropy may become dominant and PMMA is squeezed out, which leads experimentally to the formation of contamination bubbles²⁵. It is difficult to model the whole squeezing process in MD simulations because of the long timescales that are required for self-cleansing (as witnessed by the necessity of thermal annealing to clean the van der Waals heterostructures); MD analysis typically allows only simulations that correspond to less than 1 ms. In our simulations over such timescales, we saw creep of heavy hydrocarbons, but not their complete removal from the simulated nanoslits.

To avoid the timescale problem, we used metadynamics algorithms⁴² to calculate the potential of mean force, which can be regarded as a spatial free-energy profile. Also, relatively small PMMA molecules with a molecular weight M of 40,000 were simulated for computational reasons, because the results are not expected to depend on M (see below). The positions of the centre of mass for such

PMMA molecules along both relevant directions (parallel (*X*) and perpendicular (*Z*) to axis of the slit) were chosen as two variables (Extended Data Fig. 5a). Parameters from OPLS forcefield⁴¹ were used to describe the interactions among constituent atoms of the PMMA–graphene system, which include bond, angle, dihedral, improper and non-bonded (electrostatics and Lennard–Jones) interactions. Parameters for non-bonded Lennard–Jones interactions were obtained using the Lorentz–Berthelot mixing rules. The PMMA molecule was first placed on top of a graphene sheet and MD simulations were run in canonical ensembles for 10 ns with a time step of 1 fs, to reach the equilibrium at 298 K. Then, metadynamics simulations were performed for at least 300 ns. All of the calculations were carried out using LAMMPS⁴³.

As an example, in Extended Data Fig. 5 we show our results for two graphene capillaries with N=4 and N=12, which correspond to channel heights of $h\approx 13.6$ Å and $h\approx 40.8$ Å, respectively. For N=4, the PMMA molecule exhibits a higher-energy state inside the capillary than outside (Extended Data Fig. 5b). Therefore, PMMA tends to be squeezed outside or, at least, move to the edge of the entrance, as also illustrated in Fig. 2d. By contrast, the free energy of PMMA is lower inside than outside the taller capillary, as shown in Extended Data Fig. 5c and Fig. 2d.

From another perspective, the height of an adsorbed polymer molecule can be estimated as $H \approx b/\delta$, where b is the Kuhn length (for PMMA, b=1.7 nm)⁴⁴ and δ describes the ratio of the adhesion energy of a monomer to its thermal energy⁴⁴. Our MD simulations for graphene yield $\delta \approx 0.3$ at 298 K and, hence, $H \approx 5.7$ nm. This is the standard estimate, which suggests that the height of adsorbed polymer clumps should not depend on their molecular weight. To verify this assumption, we carried out MD simulations for PMMA on graphene with M=10,000-200,000. The apparent height H of the clumps was then calculated as an average over 10 ns. The results (Extended Data Fig. 6) confirm that the H values of adsorbed PMMA remain practically constant for $M \geq 40,000$. The value of $H \approx 4.0$ nm that we found is smaller but in reasonable agreement with estimate from theory (above). For lighter PMMA molecules, H tends to decrease to 2–3 nm, which is not surprising because the statistical chain model is expected to be valid only for long polymers.

The values of H that we obtained suggest that PMMA contaminants could be squeezed out of our 2D channels with h smaller than about 4.0 nm ($N \approx 12$), in (only) qualitative agreement with the experiment and simulations in Extended Data Fig. 5; the apparent height of polymer clumps may be not the best parameter to describe self-cleansing. Another possible measure of the height of PMMA molecules is their radius of gyration $R_{\rm g}$. Its perpendicular component $R_{\rm g\perp}$ provides a sense of height and so we can define the gyration height as $H_{\rm g} = 2R_{\rm g\perp}$. For self-cleansing, this parameter seems more meaningful than H because $R_{\rm g\perp}$ refers to the size of a polymer coil⁴⁴ and therefore implies a direct connection to configuration entropy. We find that $H_{\rm g}$ for PMMA contamination is about 1.5–2 nm for large M (Extended Data Fig. 6). This value matches closely the results of Extended Data Fig. 5, and our experimental data indicate the onset of self-cleansing at similar h (Fig. 2d).

Entry effects. The Knudsen flow rate through a channel with cross-section $w \times h$ is given by equation (1). For 2D channels ($h \ll w < L$) and in the Knudsen regime, α is 6

$$\alpha = \frac{h}{L} \left\{ \frac{\ln[h/w + \sqrt{1 + (h/w)^{2}}]}{h/w} + \ln\left[\frac{1 + \sqrt{1 + (h/w)^{2}}}{h/w}\right] + \frac{1 + (h/w)^{3} - [1 + (h/w)^{2}]^{3/2}}{3(h/w)^{2}} \right\}$$
(3)

which can be approximated as⁶

$$\alpha \approx \frac{16}{3\pi^{3/2}} \frac{h}{L} \ln \left(\frac{4w}{h} + \frac{3h}{4w} \right) \approx 0.958 \frac{h}{L} \ln \left(\frac{4w}{h} + \frac{3h}{4w} \right) \tag{4}$$

This expression can be simplified further to equation (2).

The Smoluchowski correction^{4–6} to equations (1), (3) and (4) can be modified to make these equations applicable also in the limit of zero f, where the factor (2-f)/f diverges, resulting in infinite Q. To this end, the flow resistance caused by the entry aperture should be taken into account and, in the first approximation¹⁰, added as a series flow resistance^{5,6}. The maximum Q is then given by equation (1) with $\alpha \equiv 1$, which correspond to perfect ballistic transport. For such transport, the enhancement coefficient is

$$K = \frac{Q(\alpha = 1)}{Q(\alpha)} \approx \frac{L}{h \ln(4w/h)}$$

shown as the dashed magenta curve in Fig. 2a. This analysis ignores the finite size d of transported gas molecules.

Now let us take into account that in our experiments h is comparable with d. In the semi-classical description (ignoring the finite de Broglie wavelength), doing so implies that molecules incident sufficiently away from the centre of the channel are unable to enter it. For incidence along the channel axis (θ =0), the effective channel width h^* is reduced from h to h-d (Extended Data Fig. 7). For molecules with a non-zero incidence angle θ , the effective entry width is reduced further to $h^*(\theta) = h - d/\cos\theta$. As an estimate, we may, for example, choose 45° as an 'average' incidence angle, which yields $h^* = h - \beta d \approx h - \sqrt{2} d$, but the coefficient β generally depends on both d and h (see below).

Taking into account the entry effect, fully ballistic transport should provide a gas flow rate of

$$Q^* = P \left(\frac{m}{2\pi RT}\right)^{1/2} w h^* \tag{5}$$

whereas the enhancement plotted in Fig. 2 (dashed curve) was calculated for $d\equiv 0$. According to equation (5), the maximum possible enhancement coefficient K^* , which takes into account both ballistic transport ($\alpha\equiv 1$) and the entry effect ($d\neq 0$) is smaller:

$$K^* = \frac{Q^*}{O(\alpha)} = \frac{h^*}{h} K = \left(1 - \frac{d}{h}\beta\right) K \tag{6}$$

To estimate β more accurately, we consider contributions from all angles θ . Incident atoms effectively 'see' only a projection of the channel opening, $h\cos\theta$. If the projected opening is smaller than d (that is, if $h^*(\theta) \leq 0$), then the atoms hit one of the edges of the aperture. Let us assume in the first approximation that all such scattered atoms are scattered away rather than guided inside the slits. This means that, when integrating, we need to exclude the trajectories that are above the critical angle $\theta_{\rm c} = \arccos(d/h)$ because they do not contribute to Q^* . This yields

$$K^* = \frac{K}{\pi} \int_{-\pi/2}^{\pi/2} 1 - \frac{d}{h \cos \theta} d\theta$$
 (7)

where the averaging is carried out over all incident angles $[-\pi/2,\,\pi/2]$ and the integrand is set to zero for $|\theta|\geq |\theta_c|.$ We can then re-write equations (6) and (7) using the average coefficient

$$\langle \beta \rangle = \frac{h}{d} \left(1 - \frac{2\theta_{\rm c}}{\pi} \right) + \frac{2}{\pi} \int_{0}^{\theta_{\rm c}} \frac{\mathrm{d}\theta}{\cos\theta}$$

Noting that

$$\int \frac{\mathrm{d}\theta}{\cos\theta} = \ln\left(\frac{1+\sin\theta}{\cos\theta}\right)$$

we obtain

$$\langle \beta \rangle = \frac{h}{d} \left(1 - \frac{2\theta_{\rm c}}{\pi} \right) + \frac{2}{\pi} \ln \left(\frac{1 + \sin \theta_{\rm c}}{\cos \theta_{\rm c}} \right)$$

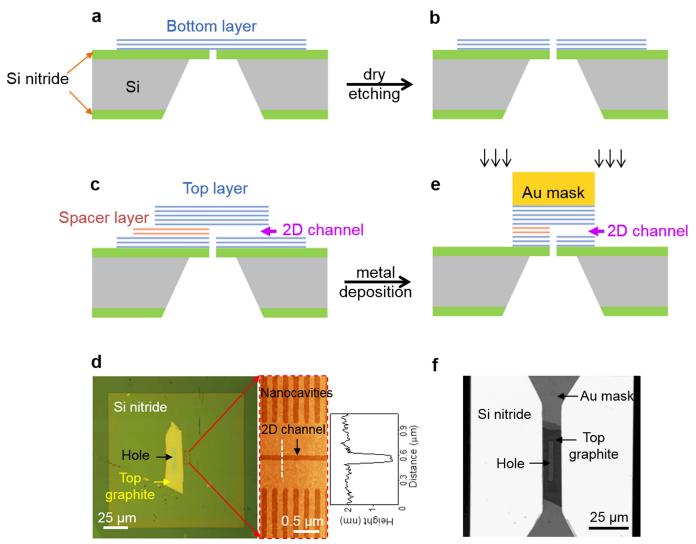
For the He-atom diameter of 2.6 Å, we find the K^* dependence shown by the solid magenta curve in Fig. 2a. Our graphene channels with N=4 exhibit the maximal enhancement, and its value agrees well with the calculated maximum possible K^* . This again indicates frictionless helium flow. The finite-size effect is expected to be enhanced by the diffraction of de Broglie waves at the entry apertures (for helium, $\lambda_{\rm B}\approx 0.5$ Å), which should lead to further reductions in K^* for the smallest N. We do not expect the quantum correction to be excessively large and so we ignore the effect in this analysis.

Data availability. The data shown in the figures and that support the findings of this study are available from the corresponding authors on reasonable request.

- Esfandiar, A. et al. Size effect in ion transport through angstrom-scale slits. Science 358, 511–513 (2017).
- Schaffer, M., Schaffer, B. & Ramasse, Q. Sample preparation for atomic-resolution STEM at low voltages by FIB. *Ultramicroscopy* 114, 62–71 (2012)
- Bentley, J. Electron density as a descriptor of thermal molecular size. J. Phys. Chem. A 104, 9630–9635 (2000).
- VandeVondele, J. et al. Quickstep: fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. Comput. Phys. Commun. 167, 103–128 (2005).
- Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* 77, 3865–3868 (1996).

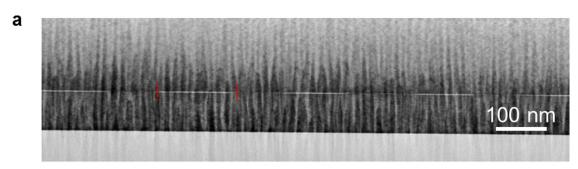


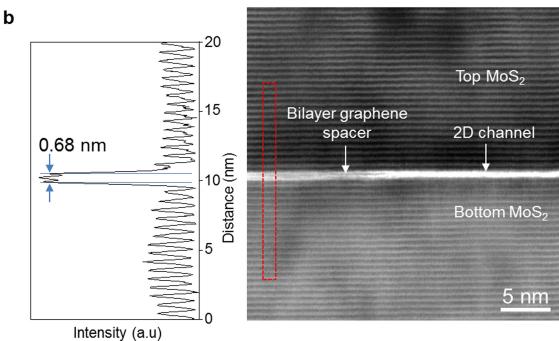
- VandeVondele, J. & Hutter, J. Gaussian basis sets for accurate calculations on mole-cular systems in gas and condensed phases. *J. Chem. Phys.* 127, 114105 (2007).
 Goedecker, S., Teter, M. & Hutter, J. Separable dual-space Gaussian pseudopotentials. *Phys. Rev. B* 54, 1703–1710 (1996).
 Lu, T. & Chen, F. Multiwfn: a multifunctional wavefunction analyzer. *J. Comput. Chem.* 23, 593 (502) (602).
- Chem. 33, 580–592 (2012).
 41. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and
- testing of the OPLS all-atom force field on conformational energetics
- and properties of organic liquids. J. Am. Chem. Soc. 118, 11225–11236
- Am. Chem. Soc. 118, 11225–11236 (1996).
 Laio, A. & Parrinello, M. Escaping free-energy minima. Proc. Natl Acad. Sci. USA 99, 12562–12566 (2002).
 Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. J. Comput. Phys. 117, 1–19 (1995).
 Rubinstein, M. & Colby, R. H. Polymer Physics Ch. 2, 3 (Oxford Univ. Press, Oxford, 2002).
- Oxford, 2003).



Extended Data Fig. 1 | **Fabrication procedures. a**, Thin crystal was transferred to cover an opening in a SiN_x membrane. **b**, The opening is extended through the bottom crystal. **c**, Spacer stripes were deposited on top of the bottom layer and etched from the back side. The top crystal is then transferred on top to fully cover the rectangular opening. **d**, Left, optical image of a single-channel capillary device made entirely from graphite; N=5. The $\mathrm{SiN}_x/\mathrm{Si}$ wafer is seen in dark green; the SiN_x membrane appears as a light-green square; the top graphite layer shows up in bright yellow; and the rectangular opening (lighter green) is indicated by the black arrow. Centre, atomic force micrograph near the channel entry, where the top graphite does not cover the spacer layer (the scan

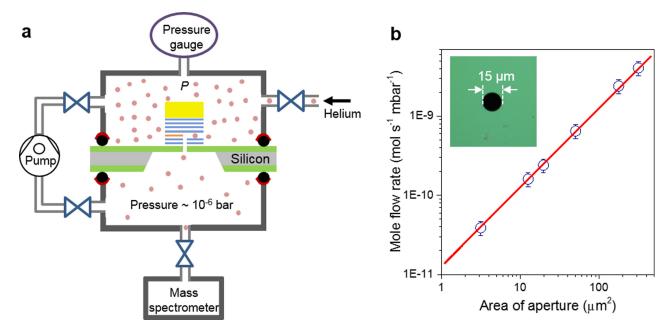
area is shown by the red contour). The height profile taken along the dotted white line is shown on the right, indicating $h\approx 1.7\pm 0.1$ nm. The side cavities perpendicular to the 2D channel were made to prevent contamination bubbles²5 across the main channel. e, A gold mask is placed on top of the trilayer assembly for final etching, to define L and to unblock the channel entry. f, Optical image of the final capillary device in the transmission mode. The SiN_x membrane is fully transparent (bright). The Au mask is partially transparent, and both the top graphite and the rectangular hole in SiN_x can be seen underneath the Au, as indicated by the arrows.





Extended Data Fig. 2 | **Visualization of 2D channels. a**, Array of 2D slits made entirely from MoS_2 , as imaged in bright-field STEM. For guidance, the edges of one of the channels are indicated by red marks. The vertical stripes are from the curtaining effect caused by ion milling³². **b**, High-magnification STEM image of a 2D channel with the top and bottom walls made from MoS_2 and bilayer graphene as the spacer (right panel). The

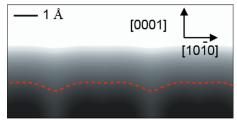
channel is white in the bright-field image, and atomic layers of MoS_2 can be seen as dark lines running parallel to the channel. The left panel shows a contrast profile across the region indicated by the red rectangle. Cross-sectional images of 2D slits made entirely from graphite crystals can be found in ref. 13 .



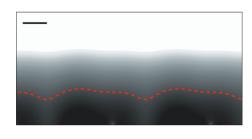
Extended Data Fig. 3 | **Helium permeation measurements. a**, Schematic of our experimental set-up. **b**, Helium flow through round apertures of various diameters as measured by our He-leak detector (symbols).

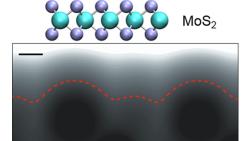
Red line, expected Knudsen flow through these apertures (no fitting parameters). Inset, optical image of one of the apertures. The error bars are from measurements using different devices.

Graphene

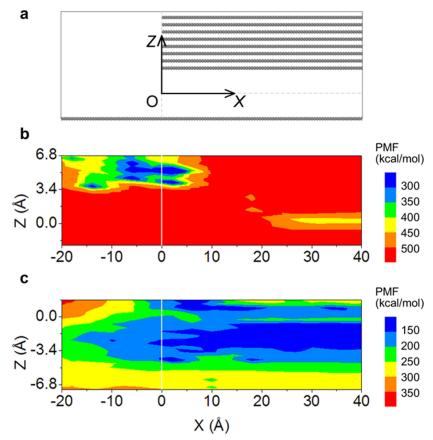






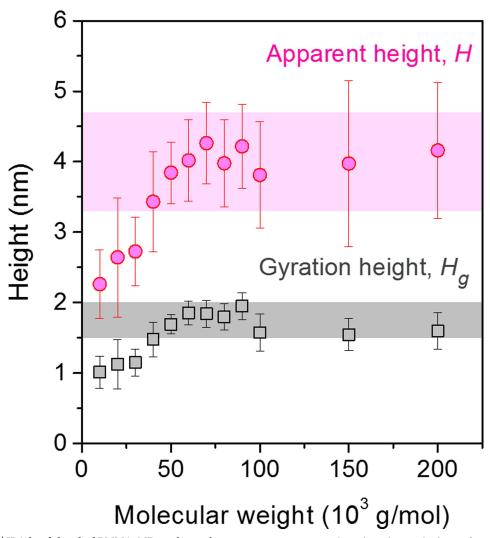


Extended Data Fig. 4 | Intrinsic roughness of atomically flat surfaces. Electron-density profiles near graphite, h-BN and MoS_2 surfaces are shown. Schematics of the atomic structures are shown on top. The red curves indicate the thermal exclusion surfaces.



Extended Data Fig. 5 | MD simulations for a heavy polymer molecule inside ångström-scale channels. a, Sketch of our simulation set-up. b, c, Energy of a PMMA molecule (M=40,000) for slits with N=4 (b) and

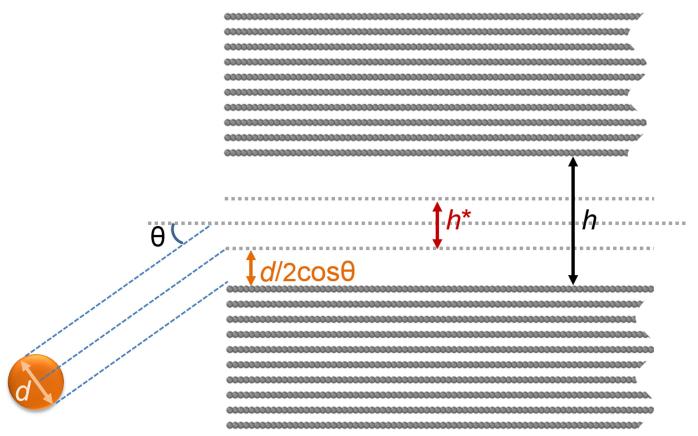
 $N\!=\!12$ (c). The axes show the position of the centre of mass with respect to the entrance edge. The origin of the axes is shown in **a**. The colour scales to the right show the relative free energy (potential of mean force, PMF).



Extended Data Fig. 6 | **Height of absorbed PMMA.** MD results are shown for the apparent (circles) and gyration (squares) heights of PMMA on graphene. The shaded areas indicate the standard errors using the data for

 $M \geq$ 40,000. Error bars show the standard error from our simulation runs lasting 10 ns.





Extended Data Fig. 7 | **Finite-size entry effect.** If an incident atom of diameter d hits one of the edges of the channel, it can be reflected. To

avoid this, the centre trajectory should be $d/(2\cos\theta)$ away from the edge, effectively reducing the entry aperture to $h^*(\theta) = h - d/\cos\theta$.



Heterointerface effects in the electrointercalation of van der Waals heterostructures

D. Kwabena Bediako^{1,6}, Mehdi Rezaee^{2,3,6}, Hyobin Yoo¹, Daniel T. Larson¹, S. Y. Frank Zhao¹, Takashi Taniguchi⁴, Kenji Watanabe⁴, Tina L. Brower-Thomas⁵, Efthimios Kaxiras^{1,3} & Philip Kim^{1,3}*

Molecular-scale manipulation of electronic and ionic charge accumulation in materials is the backbone of electrochemical energy storage¹⁻⁴. Layered van der Waals (vdW) crystals are a diverse family of materials into which mobile ions can electrochemically intercalate into the interlamellar gaps of the host atomic lattice^{5,6}. The structural diversity of such materials enables the interfacial properties of composites to be optimized to improve ion intercalation for energy storage and electronic devices⁷⁻¹². However, the ability of heterolayers to modify intercalation reactions, and their role at the atomic level, are yet to be elucidated. Here we demonstrate the electrointercalation of lithium at the level of individual atomic interfaces of dissimilar vdW layers. Electrochemical devices based on vdW heterostructures¹³ of stacked hexagonal boron nitride, graphene and molybdenum dichalcogenide (MoX₂; X = S, Se) layers are constructed. We use transmission electron microscopy, in situ magnetoresistance and optical spectroscopy techniques, as well as low-temperature quantum magneto-oscillation measurements and ab initio calculations, to resolve the intermediate stages of lithium intercalation at heterointerfaces. The formation of vdW heterointerfaces between graphene and MoX2 results in a more than tenfold greater accumulation of charge in MoX2 when compared to MoX₂/MoX₂ homointerfaces, while enforcing a more negative intercalation potential than that of bulk MoX₂ by at least 0.5 V. Beyond energy storage, our combined experimental and computational methodology for manipulating and characterizing the electrochemical behaviour of layered systems opens new pathways to control the charge density in two-dimensional electronic and optoelectronic devices.

To examine the role of the vdW heterointerface in intercalation, we assembled layers of graphene, molybdenum dichalcogenides (MoX₂, X = S, Se) and hexagonal boron nitride (h-BN) into various precise arrangements. Figure 1a shows a series of five different heterostructures (structures I to V) created using vdW assembly 14. Structure I is a simple vdW structure of graphene encapsulated by h-BN; this structure was the subject of our previous studies¹⁵ and serves as a reference point in the present study. Structures II-V are combinations of atomically thin single crystals of graphene and MoX2 encapsulated by h-BN, with several vdW heterointerfaces between atomic layers. The etched boundaries of the vdW stacks are exposed to the electrolyte. These electrochemical device architectures were investigated as the working electrodes of on-chip microelectrochemical cells, as shown in Fig. 1b, c. Using the recently developed Hall potentiometry method (see Methods), we can extract both the longitudinal resistance, R_{xx} , of the heterostructure working electrode as well as the Hall carrier density, $n_{\rm H}$, while the reaction driving force (potential, E) is altered ¹⁵. As such, the progress of the electrochemical reaction can be precisely monitored in this mesoscopic system.

Figure 2 presents an exemplary set of results for electrointercalation of a heterostructure stack of structure II (h-BN/MoS₂/graphene/h-BN). From the behaviour of R_{xx} and n_H as a function of E, four distinct

phases (phase 1-4) can be distinguished in the electrochemical reaction. This in situ monitoring of R_{xx} and n_H provides more direct information regarding the stages of intercalation than do the traditional electrochemical approaches because it is insensitive to extraneous interfacial reactions (see Extended Data Fig. 1 for a comparison). The transport features in phase 1 (for E > -2.3 V) replicate the purely electrostatic doping behaviour observed in electric double-layer gating of graphene¹⁶. With increasingly negative E, several intercalation processes occur, as evidenced by pronounced jumps in R_{xx} and n_H . The latter features in phases 3 and 4, specifically the peak in R_{xx} that occurs together with the surge in $n_{\rm H}$, are key signatures of ion intercalation involving a high-mobility graphene layer 15. The intercalation process results in a decline in electron mobility, as Li⁺ ions become closely associated with the graphene lattice and act as scattering sites for mobile electrons¹⁷. Ultimately, the resistance decreases as mounting carrier densities supersede this sudden decrease in mobility¹⁵. Deintercalation by sweeping E towards a potential of 0 V reverses doping and recovers R_{xx} and n_H to values similar to those of the pristine heterostructure (Extended Data Fig. 1a). The total carrier densities for intercalated structure-II stacks approach $n_{\rm H} = 2 \times 10^{14} \, {\rm cm}^{-2}$ (Extended Data Fig. 2), which is between three and ten times the maximal densities observed¹⁵ for intercalated structure I ((2–7) \times 10¹³ cm⁻²).

Insight into the participation of MoX₂ in this electrochemical reaction was provided by operando photoluminescence and Raman spectroelectrochemistry. In Fig. 2b, photoluminescence data reveal distinct changes in the optical profile of the semiconducting 1H-MoS₂ layer. Specifically, we found that the photoluminescence peak that is consistent with the formation of negatively charged trions (A⁻)¹⁸ appears at the later stage of the intercalation process (E < -3 V), which indicates that a strongly electron-doped 1H-MoS₂ phase persists immediately before the main intercalation stage. Beyond this point the photoluminescence is fully quenched, and Raman spectral features of the MoS₂ (and graphene) layer are lost owing to Pauli blocking 15,19,20 (Extended Data Fig. 3a). Deintercalation recovered the original Raman signatures of graphene, and revealed a reduced spectral intensity of the E_{2g}^1 and A_{1g} modes of MoS₂ as well as the emergence of a series of weak peaks between around 150 cm⁻¹ and 230 cm⁻¹ (Fig. 2c). These results are consistent with an intercalation-induced phase transition from the semiconducting H-MoS₂ phase to a metallic T phase with an additional lattice distortion (denoted as T')^{21,22}. The low-wavenumber Raman features are characteristic of the so-called 'J' modes of T-MoS2 and T'-MoS₂^{21,23-25}. Figure 2d-g and Extended Data Fig. 3c-g show the corresponding photoluminescence and Raman spectra homogeneously distributed across the interfacial areas, indicating homogeneity of the intercalation-deintercalation processes.

Low-temperature magnetotransport studies in the intercalated vdW heterostructure devices provide a new route to investigate the distribution of charge on each two-dimensional layer after intercalation. For an intercalated structure-II device, Fig. 3a shows that the Hall resistance, R_{xy} , is linear in magnetic field B, from which we estimate n_H to be

¹Department of Physics, Harvard University, Cambridge, MA, USA. ²Department of Electrical Engineering, Howard University, Washington, DC, USA. ³School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. ⁴National Institute for Materials Science, Tsukuba, Japan. ⁵Department of Chemical Engineering, Howard University, Washington, DC, USA. ⁶These authors contributed equally: D. Kwabena Bediako, Mehdi Rezaee *e-mail: pkim@physics.harvard.edu

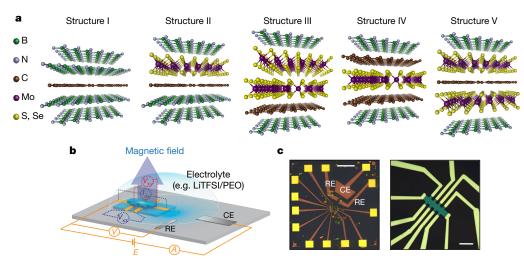


Fig. 1 | Van der Waals heterostructures for lithium intercalation.
a, Atomic models of the heterostructure series used for investigating the effects of heterolayers on intercalation capacities and thermodynamics.
b, Schematic of the mesoscopic electrochemical cell. c, Optical

micrographs of an on-chip electrochemical cell for charge transport and optical measurements during electrointercalation. Scale bars, left, 500 μm ; right, 10 μm . CE, counter electrode; RE, reference electrode.

 1.0×10^{14} cm $^{-2}$. The magnetoresistance, $R_{\rm xx}(B)$, exhibits a pronounced peak near B=0, which is presumably related to the weak localization behaviour owing to intervalley scattering of intercalated Li $^+$ ions 17 . We observe well-defined Shubnikov–de Haas (SdH) oscillations 26,27 when B>3 T, which indicates both homogeneity of the lithium-intercalated heterostructure and a high-quality two-dimensional electron gas (2DEG) with an associated carrier density of 2×10^{13} cm $^{-2}$ (see Methods). The discrepancy between $n_{\rm SdH}$ and $n_{\rm H}$ is in stark contrast with those observed for structure I (Extended Data Fig. 4), and is consistent with a two-channel electronic system, in which a higher-mobility 2DEG produces SdH oscillations corresponding to a lower-density $n_{\rm SdH} < n_{\rm H}$, while another channel contains the vast majority of electron density ($n_{\rm H} - n_{\rm SdH}$) with a lower mobility.

The decrease in amplitude of the SdH oscillation with increasing temperature (Fig. 3b) reveals an effective mass (m^*) of electrons equal to $0.11m_0$ (m_0 is the electron rest mass), close to the value of $0.099m_0$ that we obtain for intercalated structure I (h-BN/graphene/h-BN)

doped to a density of approximately 2×10^{13} cm $^{-2}$ in graphene (additional transport quantities are summarized in Extended Data Table 1). From the Landau fan diagram of structure II (Fig. 3c) we observe that the SdH quantum oscillations are strongly dependent on the voltage applied to the silicon backgate, $V_{\rm g}$, pointing to the graphene as the origin of the magneto-oscillations (see Methods). Correspondingly, we find that $n_{\rm SdH}$ and $n_{\rm H}$ exhibit the same dependence on $V_{\rm g}$ (Fig. 3d), consistent with the bottom graphene layer (around 10^{13} electrons per cm 2) serving to shield the overlying MoS $_2$ sheet (around 10^{14} electrons per cm 2) from the electrostatic influence of $V_{\rm g}$. In this case, the dependence of the total density, given by $n_{\rm H}$, simply follows the dependence of one of its components $n_{\rm SdH}$.

Having established that the carrier density distribution on these heterostructures lies strongly on the metal dichalcogenide layer, it is notable that intercalation into structure-III stacks (in which $n_{\rm H} = (1.4-1.9) \times 10^{14} \, {\rm cm}^{-2}$, see Extended Data Fig. 5) does not lead to carrier densities in excess of those seen in typical structure-II samples.

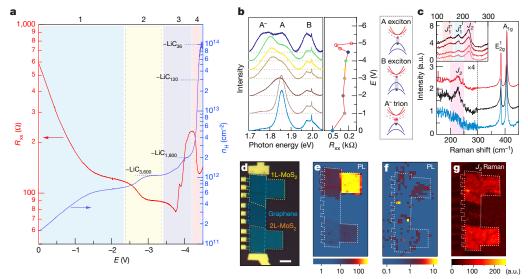


Fig. 2 | **Intercalation of structure-II devices. a**, Hall potentiogram recorded at 325 K for a graphene/MoSe₂ device (B = 0.5 T). **b**, Operando photoluminescence (left) and resistance (middle) measurements for graphene/MoS₂ at 325 K, with schematic representations of the exciton quasiparticles (right). c, Ex situ Raman spectra of a pristine (bottom), cycled (middle) and subsequently annealed (top) heterostructure. Inset,

Raman spectra after annealing of (from bottom to top) MoS₂, graphene/ MoS₂, 2L-MoS₂ and graphene/2L-MoS₂. **d**, Optical micrograph of the device used in spectroscopic mapping. Scale bar, 5 μ m. **e**-**g**, Ex situ photoluminescence (**e**, **f**) and 200–250 cm⁻¹ Raman (**g**) spatial maps of the device in **d** before intercalation (**e**) after one cycle (**f**) and after subsequent annealing (**g**).

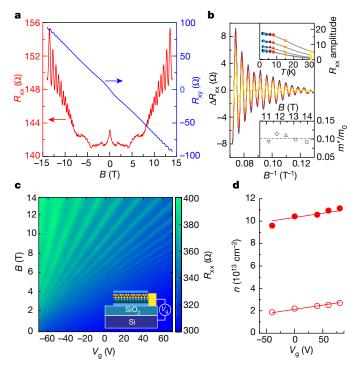


Fig. 3 | **Quantum transport. a**, Four-terminal R_{xx} and R_{xy} as a function of perpendicular B for intercalated structure II. **b**, Temperature dependence of SdH oscillations. Top inset, SdH amplitude as a function of T at five values of B. The solid lines depict fits according to the Lifshitz–Kosevich formalism. Bottom inset, effective masses m^* extracted from fits (m_0 , free electron mass). **c**, Landau fan diagram of R_{xx} (V_g , B) after intercalation. Inset, schematic of the intercalated heterostructure used with the graphene layer beneath MoS₂. **d**, Dependence of n_{SdH} (open circles) and n_{Hall} (filled circles) on V_g . Lines represent fits assuming a Si backgate capacitance of 1.2×10^{-8} F cm⁻².

This suggests that it is the graphene/MoX₂ heterointerface that contains the vast majority of intercalated ions, rather than the h-BN/MoX₂ or MoX₂/MoX₂ interfaces. To investigate this further, we created heteroarchitectures in which we designed in-plane variations of the structure type along a single graphene monolayer, as depicted in Fig. 4a. Simultaneous measurement of the transport characteristics at different lateral sections of the heterostructure devices during electrochemical polarization (Fig. 4b, Extended Data Figs. 6, 7) revealed that the onset of intercalation into graphene/MoX₂ takes place at about $\Delta E^{\circ} = +0.5$ to +0.75 V versus that of graphene/h-BN. Notwithstanding the considerably negative potential, it is noteworthy that the dichalcogenides in these graphene/MoX₂ heterostructures are not decomposed to lithium polychalcogenides as occurs in the bulk²⁰ (Extended Data Fig. 8), which indicates a widened window of electrochemical stability. This enhanced stability may be a result of dimensional confinement that restricts polysulfide/Mo⁰ nucleation and product diffusion¹². Hall resistance measurements at specific regions of the device (Fig. 4b, bottom left) unequivocally demonstrate the critical role of direct graphene/MoX₂ heterointerfaces in markedly enhancing the carrier and charge capacities in vdW heterostructure electrodes. We found that encapsulating a graphene monolayer between layers of MoX2 (as in structures V and V*), thereby creating two graphene/dichalcogenide heterointerfaces, produced intercalation capacities that were more than double those of the 'isomeric' structure-III region within the same device (Fig. 4b, bottom right). The intercalation onset potentials of the different structures (Fig. 4c and Extended Data Fig. 9a) emphasize that graphene/MoX₂ interfaces have the dominant effect on the intercalation properties; the onset of intercalation is effectively identical across structures II to V, and lies distinctly between those of structure I and bulk MoX₂. Capacities as high as 6.2×10^{14} cm⁻² (Fig. 4c, inset and Extended Data Fig. 9b, c) are attainable in structure-V devices. However, in all these structures,

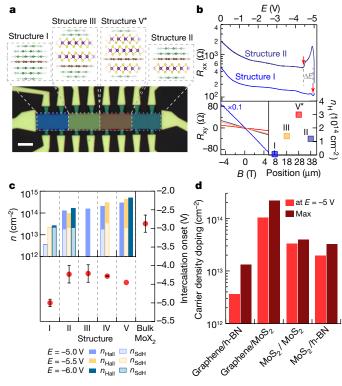


Fig. 4 | **Tuning intercalation with van der Waals heterolayers. a**, Optical micrograph (false colour) of a device consisting of several heterostructure types (depicted in the associated illustration) arrayed along a graphene monolayer. Scale bar, 5 μm. **b**, Top, R_{xx} during electrochemical gating of two regions of the device in **a**. Bottom left, Hall data following polarization of the device in **a** to -5.0 V. Bottom right, $n_{\rm H}$ from each region of the device. **c**, Intercalation onset potentials of vdW heterostructures and bulk MoX₂ (see also Extended Data Fig. 9a). Error bars, where present, represent the standard deviation (from left to right, n = 3, 6, 4, 2, 1, 3) of measurements from multiple devices and/or distinct contact pairs. Inset, mean charge densities after intercalation. $n_{\rm H}$, indicative of the total density, is depicted by the overall bar height, and graphene partial densities from SdH data (where available) are indicated by the lighter sub-bars. **d**, Estimated doping level of vdW interfaces.

the graphene density (n_{SdH}) exhibits a maximum value of around $2\times 10^{13}~\rm cm^{-2}$, which is indicative of a strong preference for charge transfer to the dichalcogenide layers (n is approximately $3\times 10^{14}~\rm cm^{-2}$ per MoX_2 layer). Assuming additive Li^+ -ion capacities, we can estimate the electrochemically accessible capacity of each vdW interface as plotted in Fig. 4d, which shows that the capacity of the graphene/ MoS_2 interface is more than ten times that of the other interfaces. These results highlight the importance of the graphene heterolayer in enhancing electrochemical charge accumulation in MoX_2 while also directing intercalation at a more negative voltage than that of bulk MoX_2 .

Finally, we explored the atomic-scale structural evolution of these layers using ex situ scanning transmission electron microscopy (STEM). As expected, data from the pristine heterostructure were fully consistent with that of H-MoS₂ (Fig. 5b). The onset of intercalation resulted in an increasingly disordered MoS₂ lattice, as evidenced by the progressive splitting of the MoS₂ Bragg spots in selected area electron diffraction (SAED) patterns. Importantly, we observed this signature of disorder at the edges of the heterostructure even before the peak in R_{xx} , while the interior remained pristine (Extended Data Fig. 10). Full intercalation resulted in the observation of a ring in the SAED data (Fig. 5c, inset). Notably, aberration-corrected STEM imaging (Fig. 5c and Extended Data Fig. 11) uncovered crystalline order within domains of approximately 5–10 nm in size (Fig. 5c, right and Extended Data Fig. 11). We also observed distinct voids of around 1 nm in size in the metal dichalcogenide layer, similar to the findings of TEM studies

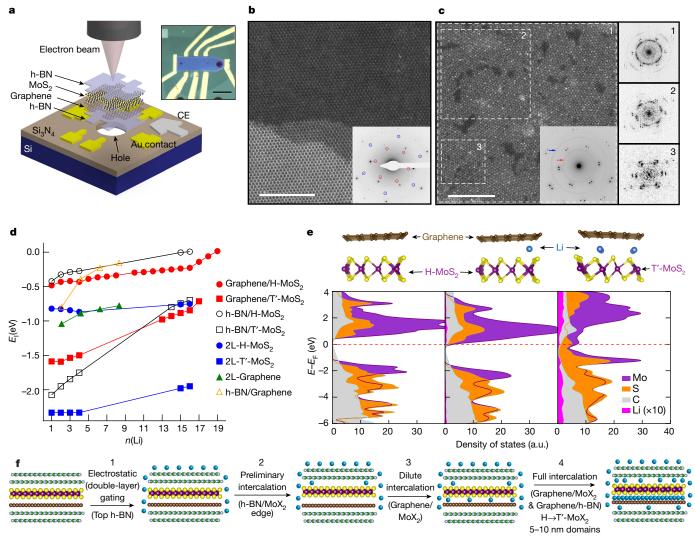


Fig. 5 | Structural evolution of van de Waals heterostructures with intercalation. a, Schematic of vdW heterostructure device for (S)TEM analysis. Inset, optical micrograph of a representative device. Scale bar, $10~\mu m.$ b, c, High-angle annular dark-field (HAADF) STEM images of structure-II devices before intercalation (b) and after one cycle (c). Scale bars, 5 nm. Insets, the corresponding SAED patterns. Diffraction features originating from $\{10\overline{1}0\}$ and $\{11\overline{2}0\}$ planes of MoS $_2$ are marked with red and blue circles or arrows, respectively. In c, fast Fourier transforms

obtained from the regions indicated with the dashed boxes are shown on the right. **d**, Computed lithium-atom binding energy as a function of the number of lithium atoms in the supercell (see Methods). 2L-graphene and h-BN/graphene data are adapted from ref. ²⁸. **e**, Computed relaxed structures (top) and density-of-states plots (bottom) for pristine (left), and lithium-intercalated (middle, 1 Li; right, 4 Li) heterostructures. **f**, Proposed mechanism of vdW heterostructure intercalation.

of chemically ("BuLi)-lithiated and exfoliated MoS_2^{22} . This structural disruption is probably caused by the strain introduced into the MoX_2 layer during lithiation and the associated progression of the H- to T' phase transformation along the lattice. Despite these structural defects, the resulting basal-plane charge transport in MoS_2 layers is reasonably high (as shown in Extended Data Table 1), which indicates that the intercalation–deintercalation process leaves the MoS_2 structure largely intact and as an electrically contiguous layer.

The tuning of intercalation potentials using vdW heterostructures is well explained by the modification of theoretical lithium binding energetics, as observed in density functional theory (DFT) calculations (Fig. 5d). First, these calculations reveal that the T'-MoS₂ phase has a considerably stronger binding affinity for lithium atoms than does H-MoS₂. As such, a local phase transformation upon doping should lead to a cooperative effect, by which it becomes increasingly favourable to intercalate lithium into that local vdW region as the dichalcogenide undergoes the semiconductor (H) to metallic (T') transformation, thereby lowering the activation barrier for Li⁺ insertion. This phase transition is manifested by the closing of the band gap, and the Fermi level crosses a band with large density of states, as shown in Fig. 5e

and Extended Data Fig. 12. Furthermore, since h-BN is an inert, widegap insulator and is non-redox-active, the energetics of initial lithium intercalation are only slightly perturbed in the case of h-BN/T'-MoS $_2$ compared to T'-MoS $_2$ /T'-MoS $_2$. By contrast, graphene heterolayers have a substantially stronger attenuating effect on the binding energy of lithium, yet still the reaction is more exergonic than that of lithium with h-BN/graphene or graphene/graphene²⁸ (Fig. 5d). As a result, we observe a positive shift in intercalation potential for the graphene/MoS $_2$ heterolayer compared to h-BN/graphene in Fig. 4b, c.

Taken together, our experimental results are consistent with the electrochemical reaction scheme presented in Fig. 5f. This mechanism involves charge transfer to both graphene and MoX_2 in the initial stages of the electrochemical gating process. Dilute concentrations of Li^+ ions are intercalated at modest potentials into MoX_2/h -BN (and MoX_2/MoX_2) heterointerfaces. However, on the basis of SAED data of our heterostructures and previous observations of slow chemical lithiation of bulk MoS_2^{20} , Li^+ ion intercalants of these interfaces appear most concentrated proximate to the heterostructure/electrolyte interface (where the electric field is strongest and some T'-Mo S_2 can be formed locally from electrostatic double-layer gating). The graphene/Mo S_2

interface possesses a unique intercalation potential that is more positive than that of graphene/h-BN, and as such this is the next interface to undergo intercalation. The exceptional electronic mobility of graphene (sufficient to display quantum oscillations even after interfacial ion intercalation) provides a lower-resistance electronic pathway—despite a lower partial carrier density—which enables its immediate interface with the MoX_2 layer to undergo ionic doping more efficiently than the adjacent MoX_2/MoX_2 homointerfaces. Eventually a highly doped, two-dimensional nanocrystalline $T^\prime\text{-MoX}_2$ is formed upon complete intercalation of the graphene/dichalcogenide heterostructure.

We note that typically, in battery electrodes consisting of layeredmaterial composites, carbonaceous additives such as graphene serve primarily to improve cyclability, particularly over the course of additional conversion reactions that can form insulating and structurally expanded conversion phases^{7,8,10,11}. These approaches do not seek to create or exploit a direct vdW contact between individual atomic layers as a means of tuning the intercalation reaction itself. Our observations of lithium-ion intercalation at individual atomic interfaces motivate the use of vdW heteroepitaxy as a promising strategy to realize new engineered functional interfaces for energy conversion and storage by manipulating the ion storage modes and 'job-sharing' characteristics of hybrid electrodes. Furthermore, our demonstrated control over intercalation energetics, the resultant spatial carrierdensity profile, and the realization of ultra-high charge densities using vdW heterointerfaces opens new possibilities for two-dimensional plasmonic device schemes²⁹ that would require large variations in charge density.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0205-0.

Received: 24 October 2017; Accepted: 26 March 2018; Published online 20 June 2018.

- Armand, M. & Tarascon, J.-M. Building better batteries. Nature 451, 652–657 (2008).
- Goodenough, J. B. & Park, K.-S. The Li-ion rechargeable battery: a perspective. J. Am. Chem. Soc. 135, 1167–1176 (2013).
- 3. Simon, P., Gogotsi, Y. & Dunn, B. Materials science. Where do batteries end and supercapacitors begin? *Science* **343**, 1210–1211 (2014).
- Maier, J. Thermodynamics of electrochemical lithium storage. Angew. Chem. Int. Ed. 52, 4998–5026 (2013).
- Ubbelohde, A. R. in Intercalated Layered Materials (ed. Lévy, F.A.) 1–32 (Riedel, Dordrecht, 1979).
- Whittingham, M. S. Electrical energy storage and intercalation chemistry. Science 192, 1126–1127 (1976).
- Pomerantseva, E. & Gogotsi, Y. Two-dimensional heterostructures for energy storage. Nat. Energy 2, 17089 (2017).
- Chhowalla, M. et al. The chemistry of two-dimensional layered transition metal dichalcogenide nanosheets. *Nat. Chem.* 5, 263–275 (2013).
 Nitta, N., Wu, F., Lee, J. T. & Yushin, G. Li-ion battery materials: present and
- future. *Mater. Today* **18**, 252–264 (2015). 10. Sun, J. et al. A phosphorene-graphene hybrid material as a high-capacity anode
- for sodium-ion batteries. *Nat. Nanotechnol.* **10**, 980–985 (2015).

 11. Oakes, L. et al. Interface strain in vertically stacked two-dimensional
- heterostructured carbon–MoS₂ nanosheets controls electrochemical reactivity. *Nat. Commun.* **7**, 11796 (2016).

 12. Zhu, C., Mu, X., van Aken, P. A., Yu, Y. & Maier, J. Single-layered ultrasmall
- Zhu, C., Mu, X., Van Aken, F. A., Yu, Y. & Maler, J. Single-layered ultrasmall nanoplates of MoS₂ embedded in carbon nanofibers with excellent electrochemical performance for lithium and sodium storage. *Angew. Chem. Int.* Ed. 53, 2152–2156 (2014).
- Geim, A. K. & Grigorieva, I. V. Van der Waals heterostructures. Nature 499, 419–425 (2013).

- Wang, L. et al. One-dimensional electrical contact to a two-dimensional material. Science 342, 614–617 (2013).
- Zhao, S. Y. F. et al. Controlled electrochemical intercalation graphene/h-BN van der Waals heterostructures. Nano Lett. 18, 460–466 (2018).
- Das, A. et. al. Monitoring dopants by Raman scattering in an electrochemically top-gated graphene transistor. Nat. Nanotechnol. 3, 210–215 (2008).
- Kühne, M. et al. Ultrafast lithium diffusion in bilayer graphene. Nat. Nanotechnol. 12, 895–900 (2017).
- Mak, K. F. et al. Tightly bound trions in monolayer MoS₂. Nat. Mater. 12, 207–211 (2013).
- 19. Malard, L. M., Pimenta, M. A., Dresselhaus, G. & Dresselhaus, M. S. Raman spectroscopy in graphene. *Phys. Rep.* **473**, 51–87 (2009).
- Xiong, F. et al. Li intercalation in MoS₂: in situ observation of its dynamics and tuning optical and electrical properties. Nano Lett. 15, 6777–6784 (2015).
- Éda, G. et al. Photoluminescence from chemically exfoliated MoS₂. Nano Lett. 11, 5111–5116 (2011).
- Eda, G. et al. Coherent atomic and electronic heterostructures of single-layer MoS₂. ACS Nano 6, 7311–7317 (2012).
- Yin, X. et al. Tunable inverted gap in monolayer quasi-metallic MoS₂ induced by strong charge-lattice coupling. Nat. Commun. 8, 486 (2017).
- Fan, X. et al. Fast and efficient preparation of exfoliated 2H MoS₂ nanosheets by sonication-assisted lithium intercalation and infrared laser-induced 1T to 2H phase reversion. Nano Lett. 15, 5956–5960 (2015).
- Singh, A. & Waghmare, U. V. in 2D Inorganic Materials Beyond Graphene (eds Rao, C. N. R. & Waghmare, U. V.) 429–431 (World Scientific, New Jersey, 2017).
- Shoenberg, D. Magnetic Oscillation in Metals (Cambridge Univ. Press, Cambridge, 1984).
- Cao, H. et al. Quantized Hall effect and Shubnikov–de Haas oscillations in highly doped Bi₂Se₃: evidence for layered transport of bulk carriers. *Phys. Rev. Lett.* 108, 216803 (2012).
- Shirodkar, S. & Kaxiras, E. Li intercalation at graphene/hexagonal boron nitride interfaces. *Phys. Rev. B* 93, 245438 (2016).
- Shirodkar, S. N. et al. Visible quantum plasmons in highly-doped few-layer graphene. Preprint at https://arxiv.org/pdf/1703.01558v1 (2017).

Acknowledgements We thank L. Jauregui, I. Fampiou and G. Kim for discussions, and S. Shirodkar for discussions and for sharing data from ref. ²⁹. The major experimental work is supported by the Science and Technology Center for Integrated Quantum Materials, National Science Foundation (NSF) grant DMR-1231319. TEM analysis was supported by Global Research Laboratory Program (2015K1A1A2033332) through the National Research Foundation of Korea. D.K.B. acknowledges partial support from the international cooperation project under the framework of the Research and Development Program of the Korea Institute of Energy Research (KIER, B8-2463-05). P.K. acknowledges partial support from the Gordon and Betty Moore Foundation's EPiQS Initiative through grant GBMF4543 and ARO MURI award W911NF14-0247. DFT calculations made use of the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University; and the Texas Advanced Computing Center at the University of Texas at Austin as part of the Extreme Science and Engineering Discovery Environment, which is supported by National Science Foundation grant ACI-1548562. K.W. and T.T. acknowledge support from the Elemental Strategy Initiative conducted by the Ministry of Education, Culture, Sports, Science and Technology, Japan and JSPS KAKENHI grant JP15K21722. Nanofabrication was performed at the Center for Nanoscale Systems at Harvard, supported in part by an NSF NNIN award ECS-00335765.

Reviewer information *Nature* thanks J. Cha and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.K.B., M.R. and H.Y. performed the experiments and analysed the data. D.K.B., S.Y.F.Z. and P.K. conceived the experiment. D.T.L. and E.K. performed the theoretical computations. K.W. and T.T. provided bulk h-BN crystals. D.K.B., M.R. and P.K. wrote the manuscript. All authors contributed to the overall scientific interpretation and edited the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0205-0.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.K. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sample fabrication. Samples were fabricated in a similar way to that described in previous work^{30,31}. In brief, mechanical exfoliation of Kish graphite (Covalent Materials Corp.) and molybdenum dichalcogenides, MoX_2 (X = S, Se) (HQ Graphene), onto p-doped silicon with 285 nm SiO₂ furnished crystals of the desired thickness, which were identified by optical contrast. Hexagonal boron nitride (h-BN) flakes of thickness 15-30 nm were similarly exfoliated and used to pick up graphene and/or MoX₂ layers in the designated order. Finally, release of these stacks onto a second flake of h-BN resulted in h-BN-encapsulated heterostructures that were subjected to annealing in high vacuum for 30 min at 350 °C. For the devices fabricated on silicon nitride membranes, thinner h-BN flakes (<5 nm) were used. Standard electron-beam lithography followed by evaporation of Cr/Pt (1 nm/9 nm) electrodes was used to define on-chip counter and pseudoreference electrodes. Reactive ion etching (RIE) using a mixture of CHF₃, Ar, and O₂ was subsequently used to shape the heterostructure into a Hall bar. Another round of lithography was used to delineate an etch mask that overlaps with the protruding legs of the Hall bar. Immediately following another RIE step, the same etch mask was used as the metal deposition mask with Cr/Pd/Au (5 nm/ 15 nm/70 nm) contacts. This resulted in a one-dimensional edge-contact to the active layers and low contact resistances.

Electrochemical doping and intercalation. In an argon-filled glove box, 3.7 ml of anhydrous acetonitrile (dried with 3 Å molecular sieves; Sigma-Aldrich) was added to 0.3 g of polyethylene oxide (PEO; Sigma-Aldrich) and 50 mg of lithium bis(trifluoromethanesulfonyl)imide (LiTFSI). After stirring overnight, a $10-15\,\mu l$ droplet of this electrolyte solution was cast onto the Si chip possessing the electrically contacted heterostructure stack, such that the droplet encompassed both the stack and the counter/reference electrodes. Rapid evaporation of the acetonitrile solvent yielded a solid polymer electrolyte for electrochemical studies. Additional extraneous solvent was removed by vacuum-drying the electrolyte overnight. Immediately before measurements the measurement device was isolated from ambient moisture and oxygen using a glass coverslip affixed to the chip carrier with vacuum grease. The device was then removed from the glove box and transferred promptly to the cryostat and vacuum-sealed.

At a temperature of 325 K, the potential between the heterostructure working electrode and Pt counter electrode was swept at a rate of approximately 1 mV s⁻¹ in the presence of a small magnetic field, B, of 0.5 T. Simultaneously, the resistance of the device was monitored by applying a small AC (17.777 Hz) current ($I_{\rm ds}$) of 0.1–1 μ A between the source and drain terminals and measuring the four-terminal longitudinal voltage drop, $V_{\rm xx}$, and Hall voltage, $V_{\rm xy}$, using a lock-in amplifier (Stanford Research SR830). The resistances $R_{\rm xx}$ and $R_{\rm xy}$ were then obtained from the equations $R_{\rm xx} = V_{\rm xx}/I_{\rm ds}$ and $R_{\rm xy} = V_{\rm xy}/I_{\rm ds}$. The Hall carrier density, $n_{\rm H}$, was then calculated from $n_{\rm H} = B/(eR_{\rm xy})$, where e is the elementary charge 1.602 \times 10⁻¹⁹ C. The Hall mobility, $\mu_{\rm H}$, during the sweep was also determined from $\mu_{\rm H} = (en_{\rm H}\rho_{\rm xx})^{-1}$, where the resistivity, $\rho_{\rm xx}$ is given by $\rho_{\rm xx} = R_{\rm xx}W/L$, where W represents the width of the conducting channel and L denotes the length of the channel between contacts. A voltmeter (Agilent 34401A Digital Multimeter) with a high internal impedance of >10 G Ω was used to measure the voltage between the heterostructure working electrode and the Pt pseudoreference electrode.

Upon reaching the desired potential, the temperature of the system was rapidly cooled to 200 K (10 K min $^{-1}$), thereby freezing the polymer electrolyte and effectively suspending any electrochemical reactions, after which additional magnetic field or temperature-dependent sweeps were conducted as desired. Further cooling to base temperature (1.8 K) was carried out at a slower rate of 2 K min $^{-1}$.

Provided that potential excursions did not exceed -6 V, we found transport behaviour to be stable to several cycles of these heterostructures.

Raman and photoluminescence spectroscopy studies. Raman and photoluminescence spectroscopy (Horiba Multiline LabRam Evolution) was conducted using 532-nm laser excitation at a power of 5–10 mW with 20-s acquisition times and four accumulations. For operando studies, the electrochemical cell or device was loaded in a glove-box environment into a cryostat (Cryo Industries of America, Inc.) with an optical window. The cryostat was then sealed, transferred out of the glove box and the measurement chamber evacuated to high vacuum for spectroelectrochemical measurements. The potential bias was swept at a rate of 2 mV s⁻¹ to the desired potentials (0, -1, -2, -3, -4 and -5 V) and held at these potentials for the acquisition of Raman and photoluminescence spectra (around 10 min) before resuming the sweep. After intercalation, the heterostructure was deintercalated by sweeping the potential to +3 V and then back to 0 V. Removal of the electrolyte was accomplished by briefly washing in deionized water followed by isopropanol. Additional spectra were subsequently acquired in this state. The deintercalated heterostructure was then annealed at 300 °C for 1 h in high vacuum.

In the pristine heterostructure, the trigonal-prismatic coordination in H-MoS $_2$ resulted in only in-plane E_{2g}^1 and out-of-plane A_{1g} modes at about 383 cm $^{-1}$ and 408 cm $^{-1}$. After a full cycle of intercalation and deintercalation, the E_{2g}^1 and A_{1g} peaks were diminished and new peaks were observed at 154, 184 and 226 cm $^{-1}$.

These low-wavenumber peaks grew in intensity with increasing numbers of MoS₂ layers—confirming their association with the dichalcogenide—and were still present, albeit slightly diminished, after annealing for 1 h at 300 °C. By contrast, the Raman peaks for the $\rm E^1_{2g}$ and $\rm A_{1g}$ modes recovered spectral intensity after annealing. The 154 and 226 cm⁻¹ peaks were attributed to the J_1 and J_2 modes of T'-MoS₂^{23,24} and the 184 cm⁻¹ feature was assigned to the J_1 mode of T-MoS₂²⁵. The corresponding Raman spectrum peak of the J_2 mode for T-MoS₂ is expected²⁵ at around 203 cm⁻¹ and therefore explains the low-wavenumber tail of the T' J_2 peak observed in Fig. 2c. We did not observe the emergence of any Raman signatures for lithium polysulfides (746 cm⁻¹)¹¹ during the entire intercalation–deintercalation processes, which suggests that the overall chemical integrity of MoS₂ remained intact upon lithiation, with a mixed phase of metastable T- and T'-MoS₂ persisting upon deintercalation, and partial recovery of H-MoS₂ after annealing.

Raman and photoluminescence spatial mapping was carried out ex situ (after removal of electrolyte) using 1.0- μ m step sizes, 2-s acquisition times and two accumulations at each pixel or step point.

Low-temperature charge transport and magnetoresistance analysis. SdH carrier densities. SdH oscillations in $R_{xx}(B)$ arise because of the formation of Landau levels at high magnetic fields²⁶. Plotting $R_{xx}(B)$ as a function of B^{-1} confirmed that these oscillations are periodic in B^{-1} with a frequency $B_{\rm F}$. The associated carrier density of the 2DEG, n_{SdH} , could then be determined from the relation $n_{\text{SdH}} = \left(\frac{geB_F}{h}\right)$, where g is the Landau level degeneracy, e is the elementary charge and h is Planck's constant. For these electron-doped graphene or MoX₂ layers, it is reasonable to assume that g = 4. Spin–valley locking in the valence band of H-MoX $_2$ layers gives rise to degeneracies of g = 2, whereas the conduction band-edges are almost spin degenerate, leading to degeneracies closer to 4 for electron-doped H-MoS₂¹⁸. Theoretical studies to date do not reveal spin-split conduction bands in T- or T'-phases of MoS₂^{32,33}. Regardless, the dependence of Hall and SdH carrier densities on the backgate voltage $V_{\rm g}$ provides additional validation for our assignment of the origin of SdH oscillations in the intercalated heterostructures. We found that, in the case of a structure-I stack consisting of a single graphene monolayer encapsulated by h-BN and biased up to E = -5.5 V for intercalation, n_{SdH} and $n_{\rm H}$ were about 2.6 \times 10¹³ cm⁻² at $V_{\rm g}$ = 0 V, changed together, and were effectively indistinguishable from each other for $V_{\rm g}$ values between $-100\,{\rm V}$ and $+100\,$ V (Extended Data Fig. 7). This reveals SdH and Hall measurements dominated by a single band as expected. In the case of the h-BN-encaspulated MoS₂/graphene heterostructure (structure II) studied here, the graphene monolayer channel is positioned in closer proximity to the backgate, underneath the MoS₂ channel. The Landau fan diagram (Fig. 3c), in which R_{xx} is plotted as a function of both V_g and B, revealed that the SdH quantum oscillations are strongly dependent on V_g , pointing to the graphene as the origin of the magneto-oscillations. Were it the case that the MoS₂ layer served as the origin of the SdH oscillations, the SdH channel would be electrostatically screened by graphene and the associated density would therefore be independent of V_g . We found that $n_{\rm SdH}$ changes with V_g in a manner consistent with the capacitance of the SiO₂/Si backgate (Fig. 3d). Indeed, we estimated the backgate capacitance, $C = 1.2 \times 10^{-8}$ F cm⁻², using $\Delta n_{\rm H} = CV_{\rm g}/e$, the value of which is in good agreement with the thickness of SiO₂ and h-BN layers serving as the gate dielectric. Considering that $n_{\rm H}$ is the total density of the heterostructure that incorporates n_{SdH} , we deduced that the density in only one layer (corresponding to $n_{\rm SdH}$) is dependent on $V_{\rm g}$. This result reveals that the layer in closest proximity to the backgate (graphene) is responsible for SdH oscillations (lower density), and therefore enables us to determine the degree of charge transfer to the individual MoX2 and graphene layers.

Effective mass determination, quantum scattering and mobilities. The effective mass, m^* , of the band giving rise to SdH oscillations was determined from the temperature dependence of the SdH amplitude, ΔR_{xx} (Fig. 4b), by fitting these data to the Lifshitz–Kosevich theory²⁷:

$$\Delta R_{xx}(B,T) \propto \frac{\frac{\alpha T}{\Delta E_{\rm N}(B)}}{\sinh\left(\frac{\alpha T}{\Delta E_{\rm N}(B)}\right)} e^{\left(-\frac{\alpha T_{\rm D}}{\Delta E_{\rm N}(B)}\right)}$$

where B is the magnetic field position of the Nth minimum in R_{xx} , $\Delta E_N(B) = heB/2\pi m^*$ is the energy gap between the Nth and (N+1)th Landau levels (m^* is the effective mass, e is the elementary charge, and h is the Planck constant), $T_D = \frac{h}{4\pi^2 \tau_q k_B}$ is the Dingle temperature (k_B is Boltzmann's constant, τ_q is

the quantum lifetime of carriers, and $\alpha=2\pi^2k_{\rm B}$ is the momentum space area including spin degeneracy. In our experiment, $\Delta E_{\rm N}$ and $T_{\rm D}$ are the only two fitting parameters. The pre-exponential in this expression is the only temperature-dependent portion and permits the straightforward determination of m^* and $\tau_{\rm q}$. In the case of intercalated structure II (h-BN/MoS₂/graphene/h-BN), we determined $m^*=0.11m_0$, and a $T_{\rm D}$ of 36.2 K, which indicates $\tau_{\rm q}=33.6$ fs and a mean free path, $l=\nu_l\tau_{\rm q}$ (where ν_l is the Fermi velocity that is taken as 10^6 m s⁻¹ for

graphene) of around 34 nm. We also determined the quantum mobility, $\mu_{\rm q} = \frac{e \tau_{\rm q}}{m^*} = 558 \, {\rm cm^2 \, V^{-1} \, s^{-1}} \, {\rm as \, compared \, to \, a \, Hall \, mobility} \, \mu_{\rm Hall} \, {\rm of \, 270 \, cm^2 \, V^{-1} \, s^{-1}}.$ These values are compared to the parameters obtained for intercalated structure I

(h-BN/graphene/h-BN) in Extended Data Table 1.

(Scanning) transmission electron microscopy. VdW heterostructures were fabricated as described above and finally transferred onto a 50-nm-thick holey amorphous silicon nitride membrane. Upon sweeping the potential, E, to the desired stage in the plot of $R_{\rm xx}$ against E, the potential was immediately returned to 0 V after which the electrolyte was removed by washing in distilled water followed by isopropanol. The delithiated heterostructure was then analysed by TEM. Abberation-corrected HAADF and bright-field STEM imaging as well as SAED were conducted on a Jeol ARM 200F equipped with a cold field-emission gun. STEM was operated at 80 kV with a probe convergence angle of 23 mrad. The inner collection semi angle for HAADF STEM imaging was 68 mrad. Bright-field and dark-field TEM imaging and SAED were performed on a Tecnai F20 operated at 120 kV. SAED data were acquired using a 300 nm aperture. Although STEM imaging is based on projected atomic structures, we still obtained atomic resolution images of mon-

olayer MoS_2 from the heterostructures by exploiting Z (atomic number)-contrast

in HAADF-STEM imaging and by using few-layer h-BN crystals. All STEM images

shown in Fig. 5b, c represent the raw, unfiltered data. For the bright-field STEM

image in Extended Data Fig. 11d, a Wiener filter³⁴ was applied to remove noise. Fast

Fourier transforms of atomic-resolution images in specified regions (Fig. 5c), as

well as the inverse fast-Fourier-transform analysis (Extended Data Fig. 11), uncov-

ers local crystallinity with domain sizes of the order of 5–10 nm. **DFT computations.** DFT computations were performed using the projector augmented wave (PAW) method³⁵ as implemented in the VASP code^{36–39}. Van der Waals interactions were included using the zero damping DFT-D3 method of Grimme⁴⁰. The heterobilayer graphene/MoS₂ system was modelled with a supercell consisting of a layer of 5×5 unit cells of fully relaxed graphene, a layer of 4×4 unit cells of MoS₂ uniformly compressed by 2.5% (in order to match the graphene lattice spacing), and over 17 Å of vacuum space between successive layers in the direction perpendicular to the layer plane. There are 98 total atoms in the bilayer supercell. All calculations were performed with an energy cut-off of 400 eV. A Γ -centred k-point mesh of $5 \times 5 \times 1$ was used for structural relaxations until all forces were smaller in magnitude than 0.01 eV (for 0 and 1 intercalated Li ions) or 0.05 eV (for 2 or more Li ions). The k-point mesh was increased to $11 \times 11 \times 1$ for electronic density of states and band structure computations. When relaxing the ions within the supercell, one Mo atom was held fixed as a reference point,

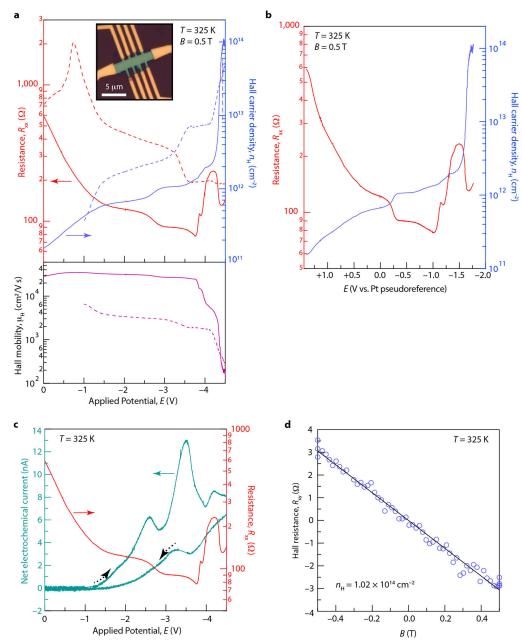
and the C atom directly above it was held fixed in the plane of the graphene layer to preserve the registration of the two layers, but was free to relax in the vertical direction. All other atoms were unconstrained. We determined the energetic stability of different intercalation states in various vdW heterostructures by calculating the binding (intercalation) energy per Li atom²⁸, $E_{\rm I}$ (Fig. 5d):

$$E_{\rm I} = \frac{1}{n} \left[E(M, n \text{Li}) - E(M) - n E(L \text{i}) \right]$$

where n is the number of Li atoms intercalated, E(M) is the energy of the empty structure M (that is, 0 Li added), E(M, nLi) is the energy of the structure M with n Li atoms intercalated and E(Li) is the energy of a Li atom in bulk lithium.

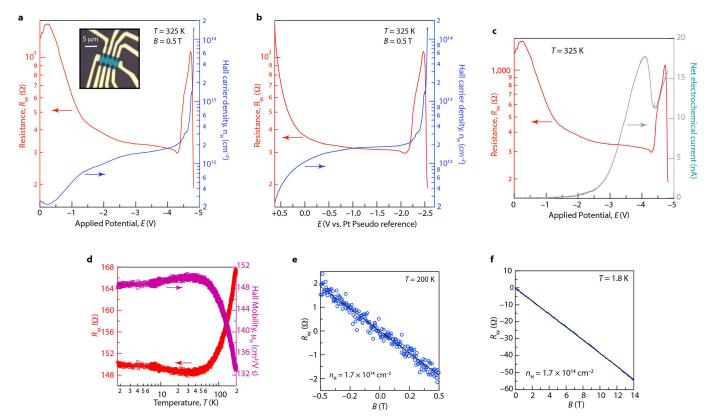
Data availability. The datasets generated and/or analysed during the current study are available from the corresponding author upon reasonable request.

- Lee, G.-H. et al. Flexible and transparent MoS₂ field-effect transistors on hexagonal boron nitride-graphene heterostructures. ACS Nano 7, 7931–7936 (2013).
- Cui, X. et al. Multi-terminal transport measurements of MoS₂ using a van der Waals heterostructure device platform. Nat. Nanotechnol. 10, 534–540 (2015).
- Kan, M. et al. Structures and phase transition of a MoS₂ monolayer. J. Phys. Chem. C 118, 1515–1522 (2014).
- Ma, F. et al. Predicting a new phase (T") of two-dimensional transition metal di-chalcogenides and strain-controlled topological phase transition. *Nanoscale* 8, 4969–4975 (2016).
- 34. Kilaas, R. Optimal and near-optimal filters in high-resolution electron microscopy. *J. Microsc.* **190**, 45–51 (1998).
- Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented wave method. *Phys. Rev. B* 59, 1758 (1999).
- Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. Phys. Rev. B 47, 558–561 (1993).
- 37. Kresse, G. & Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal-amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49**, 14251–14269 (1994).
- Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* 6, 15–50 (1996).
- Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* 54, 11169–11186 (1996).
- Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. J. Chem. Phys. 132, 154104 (2010).



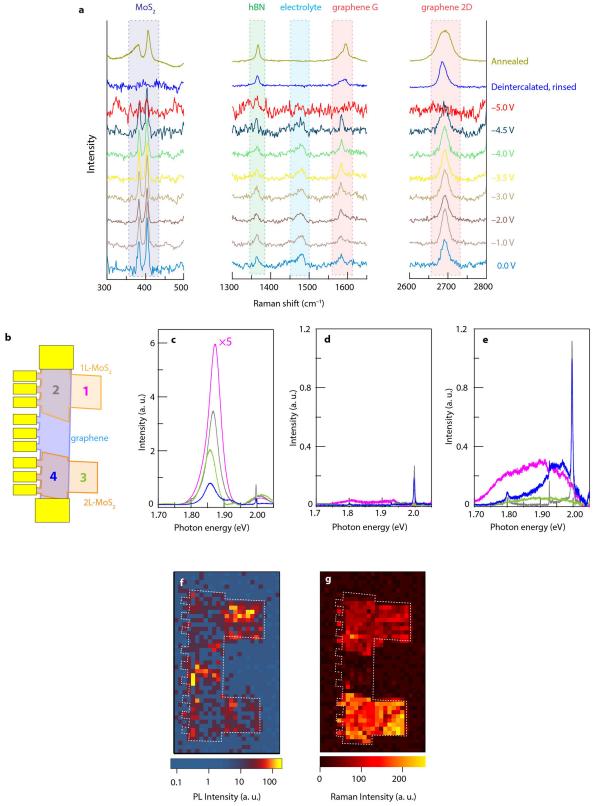
Extended Data Fig. 1 | Additional electrochemical and Hall data for the structure-II graphene/MoSe₂ stack. a, Forward (solid lines) and reverse (dashed lines) sweeps of four-probe resistance (red), Hall carrier density (blue), and Hall mobility (purple) as a function of potential at the heterostructure (versus the counter electrode/electrolyte gate—that is, in a two-electrode electrochemical configuration) in a LiTFSI/PEO electrolyte at 325 K in the presence of a magnetic field, B = 0.5 T. Inset, optical micrograph of heterostructure stack working electrode. b, Identical experiment to that in a with the resistance (red) and Hall

carrier density (blue) plotted as a function of the potential measured relative to a Pt pseudoreference electrode. **c**, Conventional cyclic-voltammetric electrochemical current response (green) overlaid with the resistance (red) over the course of the sweep, showing peaks that are difficult to assign directly to any specific reaction, probably incorporating side reactions at the Pt/electrolyte and Au/electrolyte interfaces. **d**, Hall resistance R_{xy} as a function of magnetic field at 325 K after intercalation ($E=-4.5~\mathrm{V}$).



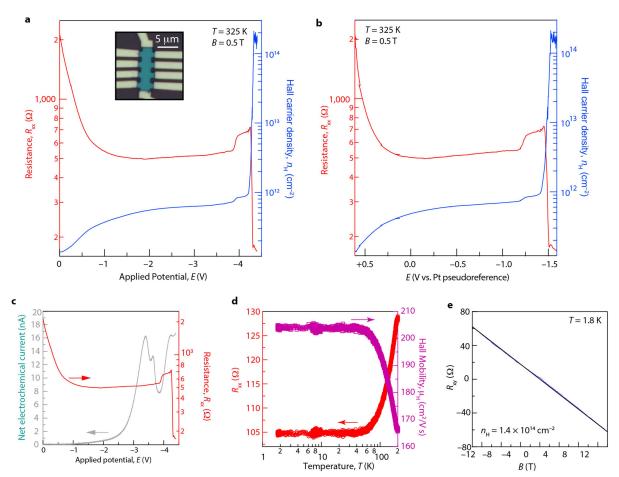
Extended Data Fig. 2 | Additional electrochemical and Hall data of the structure-II graphene/MoS $_2$ stack. a, b, Resistance (red) and Hall carrier density (blue) as a function of potential in a two-electrode (potential versus counter; a) and three-electrode (potential versus Pt pseudoreference; b) electrochemical configuration in a LiTFSI/PEO electrolyte at 325 K in the presence of a magnetic field, B, of 0.5 T. Inset, optical micrograph of heterostructure stack 'working electrode'.

c, Conventional cyclic-voltammetric electrochemical current response (grey) overlaid with the resistance (red) over the course of the sweep. **d**, Temperature dependence of resistance (red) and Hall mobility (purple) between 200 K and 1.8 K. **e**, Hall resistance, $R_{\rm xy}$, as a function of magnetic field after cooling to 200 K immediately after the termination of a sweep to -4.8 V. **f**, Hall resistance $R_{\rm xy}$ as a function of magnetic field at 1.8 K.



Extended Data Fig. 3 | Additional Raman and photoluminescence spectroscopy data. a, Raman spectra of an h-BN/graphene/MoS $_2$ structure-II device (identical device to that in Fig. 2b) over the course of electrochemical intercalation, showing the disappearance of spectral features of graphene and MoS $_2$ after full intercalation at -5.0 V, consistent with Pauli blocking in addition to the $2H \rightarrow 1T'$ phase transition of MoS $_2$. Deintercalation restores graphene peaks, and annealing at 300 °C for 1 h restores the 2H-MoS $_2$ peaks. Each spectrum is offset for clarity. b-g, Schematic diagram (b), photoluminescence spectra (c-e), photoluminescence map (f) and Raman map over the 350–450 cm $^{-1}$

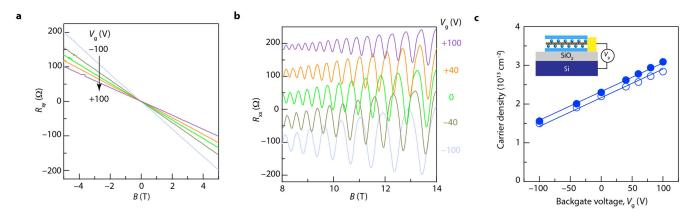
range (**g**) of an h-BN-encapsulated multi-structure device (identical device to that in Fig. 2d–g) that consists of a graphene monolayer straddling a monolayer MoS_2 crystal at one end and a bilayer MoS_2 crystal at the other. Data were acquired on the pristine stack before intercalation (**c**), after deintercalation followed by removal of electrolyte (**d**) and after subsequent annealing at 300 °C for 1 h (**e**-**g**). The sharp peak at almost 2 eV is the graphene two-dimensional (Raman scattering) peak. Photoluminescence spatial maps in the pristine state and after deintercalation are presented in Fig. 2e, f and the map of the spatial intensity of the J_2 Raman peak of the T' phase (around 226 cm $^{-1}$) after annealing is shown in Fig. 2g.



Extended Data Fig. 4 | Electrochemical and Hall data of structure III graphene/MoS₂ stack. a, Resistance (red) and Hall carrier density (blue) as a function of potential in a two-electrode (potential versus counter; a) and three-electrode (potential versus Pt pseudoreference; b) electrochemical configuration in a LiTFSI/PEO electrolyte at 325 K in the presence of a magnetic field *B* of 0.5 T. Inset, optical micrograph of heterostructure stack 'working electrode'. c, Conventional cyclic-

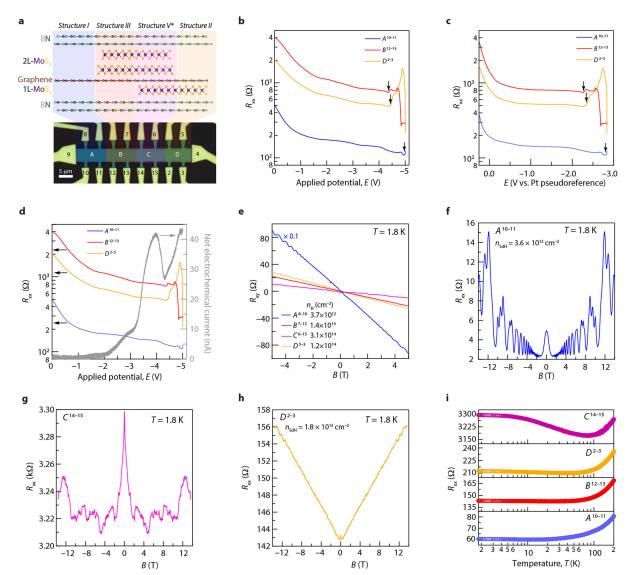
voltammetric electrochemical current response (grey) overlaid with the resistance (red) over the course of the sweep. **d**, Temperature dependence of resistance (red) and Hall mobility (purple) between 200 K and 1.8 K. **e**, Hall resistance R_{xy} as a function of magnetic field at 1.8 K. This device shows a carrier density of 1.4×10^{14} cm⁻². Maximum carrier density observed for structure-III devices is 1.9×10^{14} cm⁻².





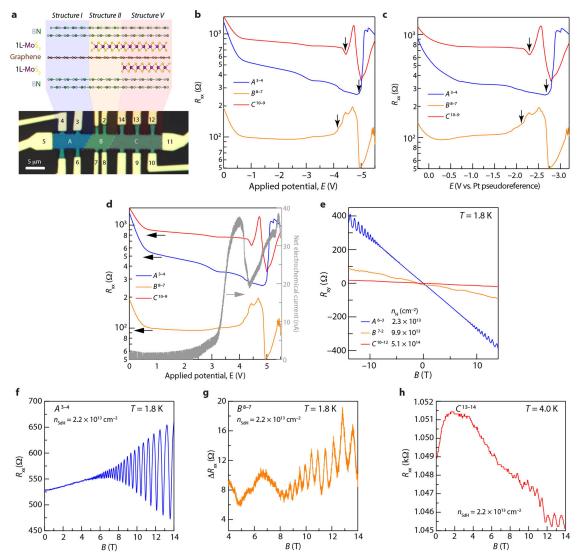
Extended Data Fig. 5 | Dependence of carrier densities of intercalated structure I on backgate voltage. a, b, Hall resistance (a) and magnetoresistance, (b; individually offset for clarity), as a function of magnetic field strength, B, in the case of a structure-I device with varying

backgate voltage, $V_{\rm g}$. c, Dependence of change in Hall (filled circles) and SdH (open circles) carrier densities on $V_{\rm g}$. Solid lines represent fits that assume a Si backgate capacitance of 1.2×10^{-8} F cm $^{-2}$.



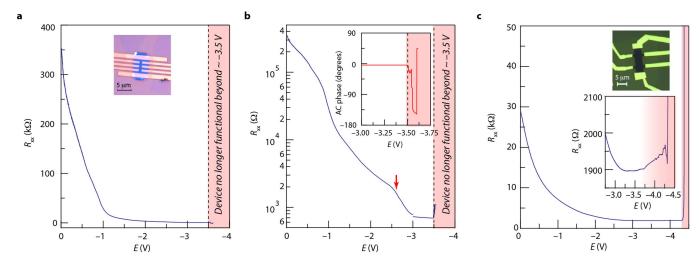
Extended Data Fig. 6 | Additional data on multi-structure-device 1. a, Optical micrograph (false colour) of a device consisting of several h-BN-encapsulated graphene/MoS $_2$ heterostructure types (depicted in the associated illustration) arrayed along a single graphene monolayer (identical device to that in Fig. 4b). b, c, Zonal resistances as a function of potential in a two-electrode (potential versus counter; b) and three-electrode (potential versus Pt pseudoreference; c) electrochemical configuration. Intercalation (indicated by the arrows) initiates at potentials approximately 0.6 V more positive at zones B (structure III) and D (structure II) than at zone A (structure I). d, Conventional cyclic-voltammetric electrochemical current response (grey) of the entire device

overlaid with the resistances of the various device regions over the course of the sweep. Cyclic voltammetry cannot distinguish between the intercalation of graphene/MoS₂ and graphene/h-BN regions in this device. **e**, Hall resistance R_{xy} as a function of magnetic field at 1.8 K for the different regions of the device after electrochemical polarization up to -5.0 V, displaying the resulting Hall carrier densities obtained. **f**-**h**, Magnetoresistance data at 1.8 K for zones A (**f**), C (**g**) and D (**h**), showing associated SdH carrier densities n_{SdH} extracted from the periodicities of oscillations in B^{-1} . **i**, Temperature dependence of resistance for the various device regions between 200 K and 1.8 K during warming.



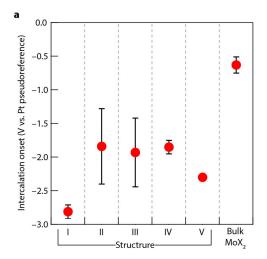
Extended Data Fig. 7 | Additional data on multi-structure-device 2. a, Optical micrograph (false colour) of a device consisting of multiple h-BN-encapsulated graphene/MoS $_2$ heterostructure types (depicted in the associated illustration) arrayed along a single graphene monolayer. b, c, Zonal resistances as a function of potential in a two-electrode (potential versus counter; b) and three-electrode (potential versus Pt pseudoreference; c) electrochemical configuration. Intercalation (indicated by the arrows) initiates at potentials approximately 0.7 V more positive at zones B (structure II) and C (structure V) than at zone A (structure I). d, Conventional cyclic-voltammetric electrochemical current

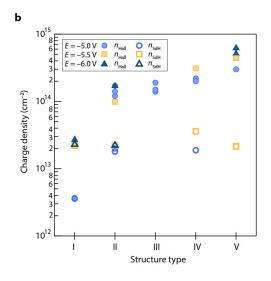
response (grey) of the entire device overlaid with the resistances of the various device regions over the course of the sweep. Cyclic voltammetry cannot distinguish between the intercalation of graphene/MoS₂ and graphene/h-BN regions in this device. **e**, Hall resistance $R_{\rm xy}$ as a function of magnetic field at 1.8 K for the different regions of the device after electrochemical polarization up to -5.5 V, displaying the resulting Hall carrier densities obtained. **f-h**, Magnetoresistance data at 1.8 K for regions A (**f**), B (**g**), and C (**h**) that reveal associated SdH carrier densities, $n_{\rm SdH}$ from the periodicities of oscillations.

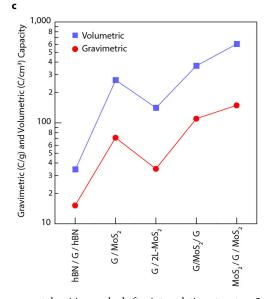


Extended Data Fig. 8 | Electrochemical gating of non-encapsulated few-layer (4–5 layers) MoX₂. a, b, Four-terminal resistance, R_{xx} , of a few-layer MoSe₂ crystal on a linear (a) and a logarithmic (b) scale, during electrochemical gating in an electrolyte comprising LiTFSI dissolved in diethylmethyl(2-methoxyethyl)ammonium TFSI (DEME-TFSI). Intercalation takes place between -2.5 V and -3 V (red arrow) and the

device loses electrical contact (demonstrated by the disruption in the phase of the lock-in amplifier (inset)) beyond –3 V. c, Four-terminal resistance, $R_{\rm xx}$, of a few-layer MoS₂ device during electrochemical gating in a LiTFSI/PEO electrolyte. As in a, the resistance of this device begins to increase at around –3.5 V and is completely insulating beyond –4.25 V, which is indicative of conversion to lithium polysulfide.

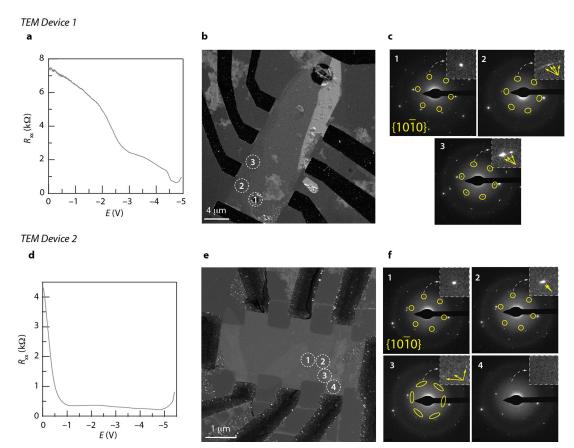






Extended Data Fig. 9 | Onset potentials and charge capacities of various heterostructures. a, Intercalation onset potentials (versus Pt pseudoreference electrode) for different vdW heterostructure types as well as few-layer MoX₂. Error bars represent standard deviations (from left to right, n = 3, 5, 4, 2, 1, 3) of measurements from multiple devices or distinct contact pairs. b, Carrier densities attained after intercalation of various h-BN/graphene/MoX₂ heterostructures. Circles, squares and triangles

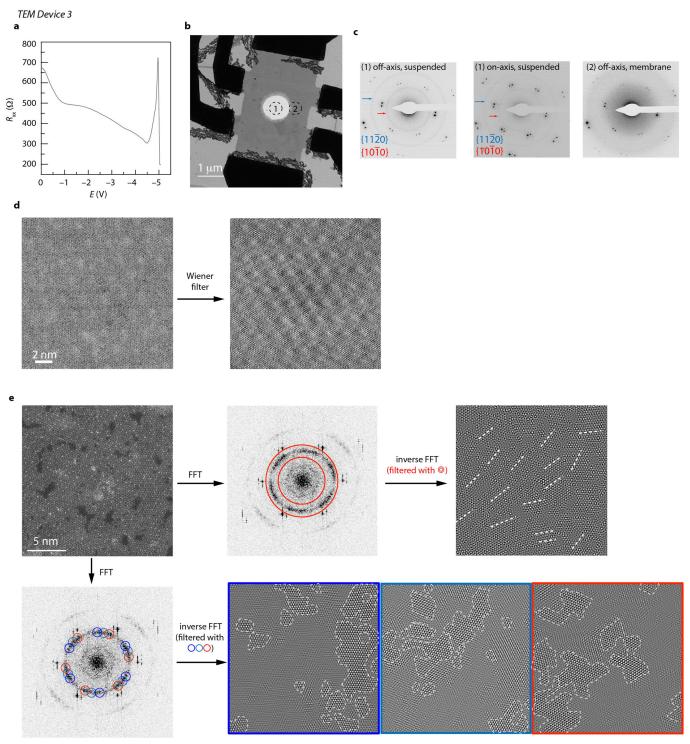
represent densities reached after intercalation at up to -5, -5.5, and -6 V, respectively. Filled symbols designate densities determined from Hall data (revealing approximate MoX_2 carrier densities, except in the case of structure I), whereas hollow symbols represent densities extracted from SdH oscillations (revealing graphene carrier densities). **c**, Average capacity values from devices in **b**, expressed in units of C g⁻¹ (gravimetric capacity) and (C cm⁻³) volumetric capacity.



Extended Data Fig. 10 | Transmission electron microscopy data of incompletely intercalated structure-II devices. a, Resistance, R_{xx} , as a function of applied potential, E, of an h-BN/MoS₂/graphene vdW heterostructure fabricated onto a 50 nm holey amorphous silicon nitride membrane. The electrochemical reaction is suspended as the increase in R_{xx} is commencing by immediately sweeping the potential back to 0 V. b, $g_{MoS_2} = 11\overline{2}0$ dark-field TEM image of the device after removal of the electrolyte. c, SAED patterns acquired from the regions designated 1, 2, and 3 in b. SAED data reveal a pristine MoS₂ structure in region 1, but splitting of the Bragg spots (insets) at the edges of the heterostructure (regions 2 and 3) indicative of the formation of two or more domains. d, Resistance, R_{xx} , as a function of applied potential, E, of an h-BN/MoS₂/

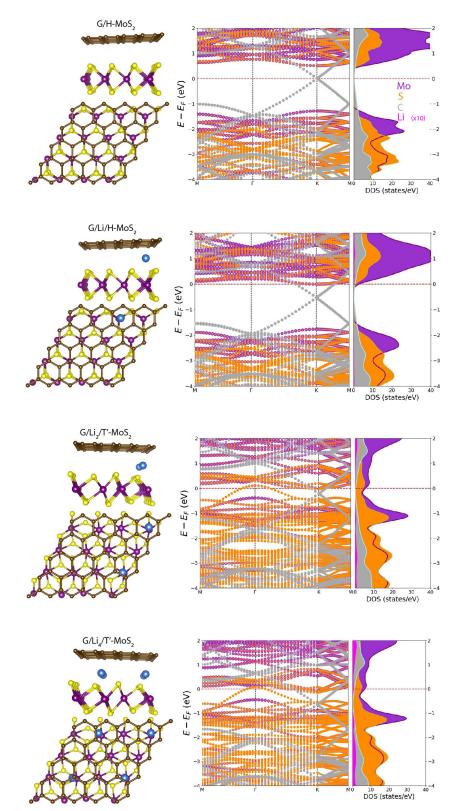
graphene/h-BN vdW heterostructure. The electrochemical reaction is suspended as $R_{\rm xx}$ approaches a maximum by immediately sweeping the potential back to 0 V. e, $g_{\rm MoS_2}=11\overline{2}0$ dark-field TEM image of the device after removal of the electrolyte. f, SAED patterns of the regions designated 1, 2, 3, and 4 in e. SAED data reveal a pristine MoS_2 structure in region 1, but strong splitting of the Bragg spots (insets) towards the edge of the heterostructure (region 3) indicative of the formation of several domains. In region 4, the diffuse scattering from the underlying amorphous silicon nitride membrane obscures any diffraction features from the MoS_2, which in that region must be considerably disordered with any domain sizes $\ll 300$ nm (the aperture size).

RESEARCH LETTER



Extended Data Fig. 11 | (Scanning) transmission electron microscopy data of the fully intercalated structure-II device. a, Resistance, $R_{\rm xx}$, as a function of applied potential, E, of an h-BN/MoS₂/graphene/h-BN vdW heterostructure fabricated onto a 50 nm holey silicon nitride membrane. The potential is reversed to 0 V after $R_{\rm xx}$ returns to a minimum (full intercalation) at around -5 V, B, Bright-field TEM image of the device after removal of the electrolyte. c, SAED patterns of the regions designated 1 and 2 in **b** in both the [0001] zone-axis (beam perpendicular to the plane of the heterostructure; middle panel) and off-zone-axis (sample tilted) conditions (left and right panels). The off-zone axis condition permits the minimization of double-diffraction phenomena associated primarily with the top and bottom h-BN flakes. SAED data at the suspended (no amorphous silicon nitride) window reveal two rings associated with

the MoS₂ layer, indicating considerable disorder in the x-y plane with a domain size \ll 300 nm (the aperture size). SAED data acquired over the membrane (region 2) cannot resolve these MoS₂ diffraction features owing to the diffuse scattering from the amorphous silicon nitride membrane in that region. **d**, Aberration-corrected bright-field STEM image of the heterostructure (left, raw data; right, filtered data), which is dominated by the h-BN in the structure. The bright periodic patches arise from the moiré pattern of the two h-BN crystals. **e**, Aberration-corrected HAADF STEM image of the device showing the nanostructure of the MoS₂ layer after one cycle. Filtered inverse fast Fourier transform (FFT) data resolve x-y rotational disorder in the MoS₂ atomic chains (top right, white dashed lines, revealing the approximate domain sizes as 5–10 nm (bottom).



Extended Data Fig. 12 | DFT-computed electronic structures of graphene/MoS₂ heterobilayers over the course of Li intercalation. Relaxed geometries (left), band structures (middle), and density-of-states plots (right) for graphene/MoS₂ structures as Li atoms are incrementally

added (top to bottom) and the phase of MoS_2 is changed from H to T'. The reason for the large carrier density in MoS_2 compared with that in graphene upon intercalation is evident from the relative density of states associated with MoS_2 compared to that of graphene.



Extended Data Table 1 \mid Charge transport parameters

$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	
	V)
$n_{\rm H}$ 2.3 × 10 ¹³ cm ⁻² 1 × 10 ¹⁴ cm ⁻²	
$n_{\rm SdH}$ 2.2 × 10 ¹³ cm ⁻² 2.0 × 10 ¹³ cm ⁻²	
m^* 0.099 m_0 0.11 m_0	
T _D 30.5 K 36.2 K	
$\tau_{\rm q}$ 39.9 fs 33.6 fs	
40 nm 34 nm	
$\mu_{\rm q}$ 712 cm ² V ⁻¹ s ⁻¹ 557 cm ² V ⁻¹ s ⁻¹	
μ_{H} 462 cm ² V ⁻¹ s ⁻¹ 270 cm ² V ⁻¹ s ⁻¹	

Comparison of transport parameters for two classes of intercalated heterostructures. The relative similarity in quantum scattering time and mean free compound support the idea that SdH oscillations observed for intercalated structure II arise from the graphene sublayer.



Extensive retreat and re-advance of the West Antarctic Ice Sheet during the Holocene

J. Kingslake^{1,6}*, R. P. Scherer^{2,6}, T. Albrecht^{3,6}, J. Coenen², R. D. Powell², R. Reese³, N. D. Stansell², S. Tulaczyk⁴, M. G. Wearing¹ & P. L. Whitehouse⁵

To predict the future contributions of the Antarctic ice sheets to sea-level rise, numerical models use reconstructions of past ice-sheet retreat after the Last Glacial Maximum to tune model parameters¹. Reconstructions of the West Antarctic Ice Sheet have assumed that it retreated progressively throughout the Holocene epoch (the past 11,500 years or so)²⁻⁴. Here we show, however, that over this period the grounding line of the West Antarctic Ice Sheet (which marks the point at which it is no longer in contact with the ground and becomes a floating ice shelf) retreated several hundred kilometres inland of today's grounding line, before isostatic rebound caused it to re-advance to its present position. Our evidence includes, first, radiocarbon dating of sediment cores recovered from beneath the ice streams of the Ross Sea sector, indicating widespread Holocene marine exposure; and second, ice-penetrating radar observations of englacial structure in the Weddell Sea sector, indicating iceshelf grounding. We explore the implications of these findings with an ice-sheet model. Modelled re-advance of the grounding line in the Holocene requires ice-shelf grounding caused by isostatic rebound. Our findings overturn the assumption of progressive retreat of the grounding line during the Holocene in West Antarctica, and corroborate previous suggestions of ice-sheet re-advance⁵. Rebound-driven stabilizing processes were apparently able to halt and reverse climate-initiated ice loss. Whether these processes can reverse present-day ice loss⁶ on millennial timescales will depend on bedrock topography and mantle viscosityparameters that are difficult to measure and to incorporate into ice-sheet models.

Recent evidence suggests that migration of the grounding line in some areas of West Antarctica during the Holocene was more complex than previously assumed^{5,7,8}. In the Weddell and Ross Sea sectors, anomalies in radar-observed englacial structure^{9,10} and isostatic rebound rates⁵ suggest that the grounding line was recently upstream of its present location. Rebound has been suggested to be a negative feedback on ice-sheet retreat¹¹⁻¹⁴ and a possible cause of grounding-line re-advance⁵, via the grounding of ice shelves^{15,16}. Better constraints on grounding-line history are important. If this history differs substantially from often-used ice-sheet reconstructions² over wide areas, then better constraints on past changes could lead to improved ice-sheet models¹ and measurements of ice-sheet mass change¹⁷. To this end, we present new evidence—from subglacial sediments and radar-observed englacial structure—for widespread re-advance of the grounding line in West Antarctica during the Holocene.

Boreholes drilled at multiple locations on the West Antarctic Ice Sheet (WAIS)—the Ross Ice Shelf Project (RISP), the Whillans Ice Stream Grounding Zone (WGZ), the Whillans Ice Stream (WIS/UpB), Subglacial Lake Whillans (SLW), the Kamb Ice Stream (KIS) and the Bindschadler Ice Stream (BIS)¹⁸—allowed recovery of subglacial sediments (till) up to 200 km inland of the present Ross Sea sector grounding line (Fig. 1). Radiocarbon analyses of 36 till samples indicate the widespread presence of young organic carbon stratigraphically

distributed through the upper metre(s) of till. The total organic carbon concentration is low, ranging from 0.2% to 0.4%, most of which is derived from Tertiary marine deposits¹⁹. Nevertheless, the organic carbon in all subglacial sediments analysed includes readily measurable radiocarbon (Extended Data Table 1).

What could be the source of this young radiocarbon? Basal melting of meteoric ice is a negligible source of radiocarbon to the subglacial environment (see Methods). Subglacial microbes cannot introduce young carbon, as they rely on legacy carbon (Methods). The samples are very unlikely to have been contaminated by modern carbon, because they were curated and sealed in different laboratories, yet yielded consistent results. Hydropotential gradients²⁰ and high basal water pressures¹⁸ drive subglacial water towards the grounding line in this region, eliminating the possibility of subglacial transport of ¹⁴C-bearing materials from the ocean to the core sites. However, observations of an active marine community just downstream of today's grounding line—more than 600 km from the open ocean (WGZ, Fig. 1; Methods)—demonstrate that radiocarbon is introduced nearly everywhere that ocean waters reach beneath the ice shelf.

We conclude that a small proportion of the organic carbon contained in the sediments was laid down under sub-ice-shelf conditions at, or upstream of, the sediment cores recently enough to allow the persistence of measurable radiocarbon. This implies that the Siple–Gould Coast grounding line was at least 200 km inland of its present position sometime after the Last Glacial Maximum (LGM). Our calculated radiocarbon ages (Extended Data Table 1) are probably much older than the most recent marine incursion, owing to dilution by more abundant radiocarbon-dead material (Methods). Moreover, ice flow transports till downstream so the grounding line may have retreated even farther inland than the core sites (Fig. 1). The proximity of radiocarbon-bearing sub-ice-stream sediments to Siple Dome (SD; Fig. 1) suggests a potential correlation with around 350 m of ice-sheet thinning during the early Holocene, documented in the Siple Dome Ice Core²¹.

On the other side of the WAIS, we conducted a 700-km-long, ground-based, ice-penetrating radar survey of Henry Ice Rise (HIR; Figs. 1, 2 and Methods). HIR is 7,000 km² in area and is grounded 310–800 m below sea level. Our survey revealed englacial structures that are inconsistent with present-day slow (less than 10 m per year) and cold-based flow conditions (see Methods).

A series of steep englacial reflectors (Fig. 2d) cluster around a basal topographic high at the northern end of HIR (Fig. 2a). These features intercept the bed, penetrate to 200–300 m above the bed, and crosscut smoothly undulating isochrones (Fig. 2d and Extended Data Fig. 1). They have similar lateral extents, orientations and spacing to extensional surface crevasses at Doake Ice Rumples (DIR; Fig. 1 and Extended Data Fig. 2). At ice rumples, ice that was floating upstream flows onto and over a bedrock high. We interpret the buried features in HIR as marine-ice-filled relic crevasses that formed when icerumple flow persisted on HIR. The crevasses were probably near-vertical while active and have been buried and deformed to varying

¹Lamont-Doherty Earth Observatory, Columbia University, New York, NY, USA. ²Department of Geology and Environmental Geosciences, Northern Illinois University, DeKalb, IL, USA. ³Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Germany. ⁴Earth and Planetary Sciences Department, University of California Santa Cruz, Santa Cruz, CA, USA. ⁵Department of Geography, Durham University, Durham, UK. ⁶These authors contributed equally: J. Kingslake, R. P. Scherer, T. Albrecht. *e-mail: jkingslake@columbia.edu

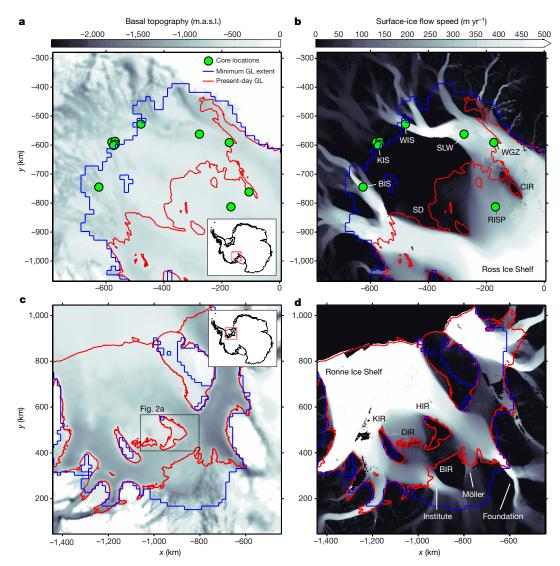


Fig. 1 | Basal topography and surface ice-flow speed in the Weddell and Ross Sea sectors of West Antarctica. a, Basal topography and bathymetry 29 and b, surface-ice flow speed 30 in the Ross Sea sector. The locations of sediment recovery are shown in green. m.a.s.l., metres above sea level. c, Basal topography and bathymetry 29 and d, surface-ice flow

speed 30 in the Weddell Sea sector. In all panels the present-day grounding line $(GL)^{31}$ is in red, while the (asynchronous) modelled minimum extent of the grounding line in each sector is in blue. Axes show polar stereographic coordinates in kilometres. Insets show locations in West Antarctica. The Institute, Möller and Foundation ice streams are labelled.

extents into steeply dipping structures by complex ice flow (Methods). Further evidence that parts of HIR were previously floating include prominent synclines in internal isochronal layers that increase in amplitude with depth, are unrelated to basal topography and truncate at the bed (Fig. 2b, c and Extended Data Fig. 1)—characteristics indicative of past ocean melting⁹.

Ice-shelf grounding on the topographic high beneath HIR—first forming ice rumples, then thickening to form the ice rise—can explain the unusual englacial structures. Contact with the ocean generates isochrone synclines where melting is focused at a static grounding line for long enough²². Ice-rumple flow generates surface crevasses similar to those observed on and downstream of DIR, which were preserved in HIR as flow stagnated. Prior to grounding, the ice shelf probably flowed approximately northward in the location of HIR. Post-grounding thickening upstream of the topographic high explains today's configuration, with the initial grounding point beneath HIR's northern extreme. An alternative interpretation is that HIR persisted throughout the Holocene and recently grew to its present size. However, we argue that complete ungrounding is more likely (see Methods). Under either scenario, we interpret a contrast in surface texture, approximately coincident with the onset of relic crevassing (Fig. 2b), as a signature of a past grounding-line configuration (Methods). The formation or

regrowth of HIR is expected to have increased the buttressing force exerted by the Ronne Ice Shelf on the upstream ice sheet, with implications for grounding-line migration and mass balance.

To explore the cause and implications of ice-rise formation (revealed by radar observations) and ice stream grounding-line retreat and readvance (revealed by radiocarbon analyses), we turned to numerical ice-sheet modelling. We simulated the post-LGM evolution of the WAIS using the Parallel Ice Sheet Model (PISM)²³ with improved descriptions of sub-shelf melting and solid Earth rebound, forced by sea-level and ice-core temperature reconstructions (see Methods). A model ensemble investigated first-order sensitivities to independent variations in parameters related to ice flow, glacial isostatic adjustment (GIA), calving, sub-shelf melting, basal traction and accumulation.

After partially compensating for uncertainty in bed topography (Methods), our simulations display remarkable agreement with the conclusions of our radiocarbon and radar analyses. Our reference simulation (Methods) demonstrates this agreement (Fig. 3 and Supplementary Video 1). In this simulation, rising sea-level and surface temperatures during the last glacial termination drive grounding-line retreat through regions currently occupied by the Ronne and Ross ice shelves. The grounding line reaches its most retreated position around 10 thousand years (kyr) before present (BP), up to approximately

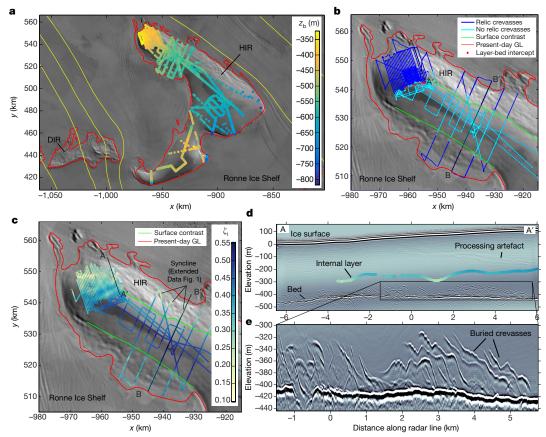


Fig. 2 | Ice-penetrating radar evidence for grounding of the Ronne Ice Shelf. a, Radar-derived ice-bed elevation beneath HIR. See Fig. 1c for location in the Weddell Sea sector. The present-day GL^{31} is in red. b, Radar lines coloured according to where relic crevasses are found. c, Normalized elevation (ζ_i) of an isochrone (Methods). Background images in \mathbf{a} - \mathbf{c} are from the MODIS mosaic of optical (red band) imagery over Antarctica (MOA), which reveals Antarctic surface morphology³².

Green lines in $\bf b$ and $\bf c$ highlight a contrast in surface texture (see Methods) that runs parallel to the present-day grounding line (GL), to the onset of relic crevassing and, on the east side, to a prominent isochrone syncline. $\bf d$, Radargram displaying examples of undulating isochrones. One isochrone is mapped using the colour map from $\bf c$. $\bf e$, Close-up view of near-bed relic crevasses with mean spacing of approximately 450 m.

300 km inland of the present-day grounding line (Fig. 3 and Extended Data Fig. 3). Retreat exposes nearly all of our core sites and the bed of HIR to the ocean. Approximately 352,000 km² of the area that is covered by grounded ice today ungrounds during retreat, resulting in lithospheric rebound of up to 175 mm per year. The rising bed eventually causes the Ross and Ronne ice shelves to ground on bathymetric highs in the locations of present-day ice rises, including HIR. Ice-rise formation increases ice-shelf buttressing, causing the grounding line to re-advance towards its present-day location (Fig. 3 and Extended Data Fig. 4; Methods). In the Amundsen Sea sector, the grounding line retreats to its modern position without substantial inland retreat and re-advance.

During this simulation, rebound-driven re-advance causes the WAIS to gain ice above the flotation level equivalent to 33 cm of sea-level fall (Weddell sector, 2 cm; Ross sector, 31 cm). Ice-volume minima in each sector are asynchronous and the minimum in whole ice-sheet volume occurs 1.5 kyr BP, at which time the ice sheet is 20 cm sea-level equivalent smaller than at present.

The timing and magnitude of the simulated grounding-line retreat and re-advance depend on model parameters, forcings, bed topography and spatial resolution (Extended Data Figs. 6 and 7; Methods). For example, increasing mantle viscosity expedites retreat, increases maximum retreat and delays re-advance. Ice-rise formation greatly enhances grounding-line re-advance and is sensitive to bed topography, which is regionally uncertain; moreover, dynamically relevant topographic features are poorly represented at the spatial resolution of the model (Extended Data Fig. 4 and Methods).

Notably, although grounding-line re-advance was not their focus, four previous Antarctic ice-sheet modelling studies—using alternative parameterizations of basal sliding, grounding-line flux and lithosphere response—also simulate Holocene grounding-line retreat and readvance in these sectors in some simulations^{24–27}.

Radiocarbon in subglacial sediments, radar-observed relic crevassing and ice-sheet modelling provide corroborating evidence that two large Antarctic catchments re-advanced to their present-day configurations during the Holocene (Fig. 3). Previous work is consistent with this conclusion, but cannot confirm or rule out Holocene retreat and re-advance (see Methods). Moreover, previous authors have found evidence for localized re-advance and suggested rebound as a cause^{5,10}. However, ice-sheet reconstructions used to tune ice-sheet models and to correct mass-balance observations do not at present include large-scale grounding-line re-advance^{1,2}. Updating these reconstructions to include re-advance could influence ice-sheet gravimetry and altimetry¹⁷, and sea-level projections. Furthermore, we hypothesize that the grounding line in the Weddell and Ross Sea sectors may be capable of retreating far inland of its present position without triggering runaway ice-sheet collapse.

We note that our model does not simulate retreat and rebound-driven re-advance in the Amundsen Sea sector (Fig. 3), where present-day retreat of the grounding line is causing concern about future runaway collapse⁶ and recent re-advance could explain observed sub-shelf iceberg ploughmarks²⁸. Our findings motivate future work to examine whether rebound-driven mechanisms could slow or reverse this retreat on millennial timescales.

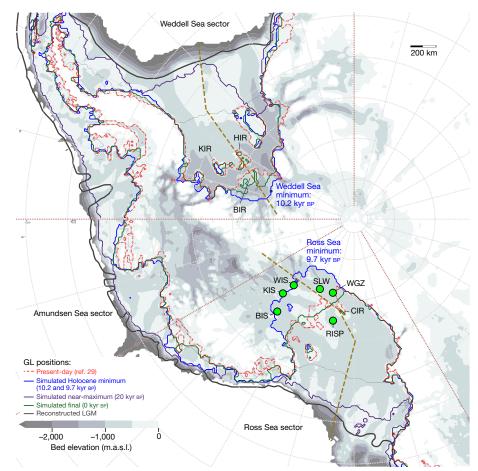


Fig. 3 | Modelled grounding-line retreat and re-advance due to lithospheric rebound. The position of the WAIS grounding line at 20 kyr BP in our reference simulation is shown in violet, with a recent LGM ice-sheet reconstruction in black (ref. ², scenario B). The ice sheet asynchronously reaches a minimal extent in the Weddell and Ross Sea sectors at 10.2 kyr BP and 9.7 kyr BP respectively (blue). The grounding

line (GL) then re-advances towards its present-day location²⁹ (red). The final simulated grounding-line position is in green. The locations of Siple–Gould Coast sediment cores and selected ice rises are indicated. Brown dashed lines show cross-sections used for Extended Data Figs. 3, 6 and 7. Red dotted lines show longitude-defined sectors. Background shading shows basal topography and bathymetry²⁹.

Rising eustatic sea levels and temperatures were major climate-related drivers of ice-sheet retreat during and after the last glacial termination. By contrast, it appears that climate-independent lithospheric rebound and ice-shelf grounding were the main drivers of grounding-line re-advance during the Holocene. The impact of rebound on the ice sheet depends sensitively on bedrock topography and mantle viscosity (see Methods). Accurate mapping of potential grounding points and improved parameterization of uplift are needed to forecast the direction and rate of future grounding-line migration in West Antarctica.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0208-x.

Received: 30 August 2017; Accepted: 28 March 2018; Published online: 13 June 2018

- DeConto, R. M. & Pollard, D. Contribution of Antarctica to past and future sea-level rise. Nature 531, 591–597 (2016).
- Bentley, M. J. et al. A community-based geological reconstruction of Antarctic lce Sheet deglaciation since the Last Glacial Maximum. Quat. Sci. Rev. 100, 1–9 (2014).
- Conway, H. et al. Past and future grounding-line retreat of the West Antarctic Ice Sheet. Science 286, 280–283 (1999).
- Spector, P. et al. Rapid early-Holocene deglaciation in the Ross Sea, Antarctica. Geophys. Res. Lett. 44, 7817–7825 (2017).

- Bradley, S. L. et al. Low post-glacial rebound rates in the Weddell Sea due to Late Holocene ice-sheet readvance. Earth Planet. Sci. Lett. 413, 79–89 (2015).
- Scambos, T. A. et al. How much, how fast? A science review and outlook for research on the instability of Antarctica's Thwaites Glacier in the 21st century. Global Planet. Change 153, 16–34 (2017).
- Goodwin, I. D. Did changes in Antarctic ice volume influence late Holocene sea-level lowering? Quat. Sci. Rev. 17, 319–332 (1998).
- Halberstadt, A. R. W., Simkins, L. M., Greenwood, S. L. & Anderson, J. B. Past ice-sheet behaviour: retreat scenarios and changing controls in the Ross Sea, Antarctica. Cryosphere 10, 1003–1020 (2006).
- Catania, G. A. et al. Evidence for floatation or near floatation in the mouth of Kamb Ice Stream, West Antarctica, prior to stagnation. J. Geophys. Res. Earth Surf. 111, F01005 (2006).
- Siegert, M. et al. Late Holocene ice-flow reconfiguration in the Weddell Sea sector of West Antarctica. Quat. Sci. Rev. 78, 98–107 (2013).
- Adhikari, S. et al. Future Antarctic bed topography and its implications for ice sheet dynamics. Solid Earth 5, 569–584 (2014).
- Gomez, N., Pollard, D. & Holland, D. Sea-level feedback lowers projections of future Antarctic Ice-Sheet mass loss. Nat. Commun. 6, 8798 (2015).
- Greischar, L. L. & Bentley, C. R. Isostatic equilibrium grounding line between the West Antarctic inland ice sheet and the Ross Ice Shelf. *Nature* 283, 651–654 (1980).
- Konrad, H. et al. Potential of the solid-Earth response for limiting long-term West Antarctic Ice Sheet retreat in a warming climate. Earth Planet. Sci. Lett. 432, 254–264 (2015).
- Matsuoka, K. et al. Antarctic ice rises and rumples: their properties and significance for ice-sheet dynamics and evolution. *Earth Sci. Rev.* 150, 724–745 (2015).
- Thomas, R. H. The creep of ice shelves: interpretation of observed behavior. J. Glaciol. 12, 55–70 (1973).
- Hanna, E. et al. Ice-sheet mass balance and climate change. Nature 498, 51–59 (2013).
- Kamb, B. in The West Antarctic Ice Sheet: Behavior and Environment (eds Alley, R. B. & Bindschadler, R. A.) 157–199 (American Geophysical Union, Washington DC, 2001).



- Scherer, R. P. Quaternary and tertiary microfossils from beneath ice stream B: evidence for a dynamic West Antarctic ice sheet history. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 90, 395–412 (1991).
- Livingstone, S. et al. Potential subglacial lake locations and meltwater drainage pathways beneath the Antarctic and Greenland ice sheets. Cryosphere 7, 1721–1740 (2013).
- Price, S. F., Conway, H. & Waddington, E. D. Evidence for late Pleistocene thinning of Siple Dome, West Antarctica. J. Geophys. Res. Earth Surf. 112, F03021 (2007).
- 22. Catania, G., Hulbe, C. & Conway, H. Grounding-line basal melt rates determined using radar-derived internal stratigraphy. *J. Glaciol.* **56**, 545–554 (2010).
- 23. Winkelmann, R. et al. The Potsdam Parallel Ice Sheet Model (PISM-PIK)—part 1: model description. Cryosphere 5, 715–726 (2011).
- Pollard, D. et al. Large ensemble modeling of the last deglacial retreat of the West Antarctic Ice Sheet: comparison of simple and advanced statistical techniques. Geosci. Model Dev. 9, 1697–1723 (2016).
- Golledge, N. R. et al. Antarctic contribution to meltwater pulse 1A from reduced Southern Ocean overturning. Nat. Commun. 5, 5107 (2014).
- Maris, M. N. A. et al. A model study of the effect of climate and sea-level change on the evolution of the Antarctic Ice Sheet from the Last Glacial Maximum to 2100. Clim. Dyn. 45, 837–851 (2015).
- Pollard, D., Gómez, N. & Deconto, R. M. Variations of the Antarctic ice sheet in a coupled ice sheet-Earth-sea level model: sensitivity to viscoelastic Earth properties. J. Geophys. Res. Earth Surf. 122, 2124–2138 (2017).
- Graham, A. G. et al. Seabed corrugations beneath an Antarctic ice shelf revealed by autonomous underwater vehicle survey: origin and implications for the history of Pine Island Glacier. J. Geophys. Res. Earth Surf. 118, 1356–1366 (2013).
- Fretwell, P. et al. Bedmap2: improved ice bed, surface and thickness datasets for Antarctica. Cryosphere 7, 375–393 (2013).
- Rignot, E., Mouginot, J. & Scheuchl, B. Ice flow of the Antarctic ice sheet. Science 333, 1427–1430 (2011).
- Depoorter, M. A. et al. Calving fluxes and basal melt rates of Antarctic ice shelves. Nature 502, 89–92 (2013).
- Haran, T. et al. MODIS mosaic of Antarctica 2003–2004 (MOA2004) image map. US Antarctic Program Data Center https://doi.org/10.7265/N5ZK5DM5 (2005).

Acknowledgements J.K. and the Weddell Sea fieldwork were funded by Natural Environmental Research Council (NERC) grant NE/J008087/1, led by R. Hindmarsh. Logistical support was provided by many members of the British Antarctic Survey's air unit and field operations team. We particularly thank I. Rudkin and S. Webster for assistance in the field. We also thank H. Pritchard for supplying bed elevation data and Schlumberger Limited for a software donation. PISM development is supported by NASA grants NNX13AM16G and NNX13AK27G. T.A. is supported by the Deutsche Forschungsgemeinschaft (DFG) in the framework of the priority program 'Antarctic Research with comparative investigations in Arctic ice areas' through grants LE1448/6-1 and LE1448/7-1.We acknowledge the European Regional Development Fund, the German Federal Ministry of Education and Research, and the Land Brandenburg for providing high-performance computer resources at the Potsdam Institute for Climate Impact Research. We also thank the Gauss Centre for Supercomputing e.V. (http://www.gauss-centre.eu) for

providing computing time on the GCS Supercomputer SuperMUC at Leibniz Supercomputing Centre (http://www.lrz.de; project code pr94ga). We thank C. Buizert for providing ice-core temperature reconstructions; D. Peltier for access to eustatic sea-level reconstructions; J. Lenaerts for surface mass balance data from the RACMO climate model; and S. Jamieson for providing the RAISED consortium's grounding-line reconstructions. R.P.S., J.C., R.D.P. and S.T. were funded by National Science Foundation (NSF) WISSARD Project grants ANT-0839107, ANT-0839142, ANT-0838947 and ANT-0839059. Collection of subglacial sediment samples at Subglacial Lake Whillans and the Whillans Grounding Zone was facilitated by the US Antarctic Program and the efforts of multiple field support teams, including the drilling team from the University of Nebraska-Lincoln and WISSARD traverse personnel, as well as by Air National Guard and Kenn Borek Air who provided air support. WIS, KIS and BIS samples were recovered by B. Kamb's program at the California Institute of Technology (1988–2001), which included R.P.S. and S.T.; samples from the US Antarctic Program's Ross Ice Shelf Project (1977–1979) cores were made available for study by the US Antarctic Sediment Core Repository, Florida State University. P.L.W. is funded by a NERC Independent Research Fellowship (NE/K009958/1). This research is a contribution to the Scientific Committee on Antarctic Research (SCAR) Solid Earth Response and Influence on Cryosphere Evolution (SERCE) program. We thank R. Arthern, R. Bell, R. Hindmarsh, C. Martín, J. Southon and K. Tinto for discussions that contributed to this study. We particularly thank D. Pollard for sharing ideas and unpublished Penn State model outputs for discussion.

Reviewer information *Nature* thanks R. Drews, J. Smith and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors contributed to manuscript preparation. T.A., J.K. and R.P.S. are co-lead authors with equal contributions; others are listed alphabetically. J.K. designed and conducted the Weddell Sea sector ice-penetrating radar survey and led the preparation of the manuscript. R.P.S., J.C., R.D.P. and S.T. collected and analysed sub-ice sediment samples as part of the WISSARD and earlier drilling projects in the Ross Sea sector. N.D.S. and J.C. prepared samples and interpreted ¹⁴C and ¹³C results. T.A. ran the PISM simulations and an extended analysis of parameter sensitivity. R.R. designed and analysed experiments for disentangling drivers of re-advance. M.G.W. analysed radar data from the Weddell Sea sector. P.L.W. provided input on parameterization of solid-Earth rebound and sea-level forcing for the model experiments.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0208-x

 $\begin{tabular}{ll} \textbf{Supplementary information} is available for this paper at https://doi.org/10.1038/s41586-018-0208-x. \end{tabular}$

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.K. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sediments. Radiocarbon and ¹³C analyses of glacial tills. Subglacial sediments have been recovered during multiple field seasons by hot-water drilling through the southern Ross Ice Shelf and grounded West Antarctic Ice Sheet. Sub-ice-shelf core samples include the Ross Ice Shelf Project (RISP, 1978; ref. ³³) and the Whillans Ice Stream Grounding Zone (WGZ, 2015), recovered as part of the Whillans Ice Stream Subglacial Access Research Drilling (WISSARD) Project. The WISSARD Project also recovered cores from beneath grounded ice at Subglacial Lake Whillans (SLW, 2013). Sub-ice-stream samples further upstream were recovered from the Whillans (WIS/UpB, 1989, 1991, 1995, SLW, 2013), Kamb (KIS, 1995, 1996, 2000) and Bindschadler (BIS, 1998) ice streams¹⁸. The sediments recovered are tills with a matrix derived in part from strata that accumulated during multiple intervals of terrestrial, coastal and open marine deposition in West Antarctica. Source strata integrated into the tills are dated by microfossils 19,34. They include terrestrial plant spores dating back to the Devonian, but are dominantly Miocene-age diatoms, reflecting the abundance of Miocene marine strata in the embayment. The youngest diatoms present are of Pleistocene age, representing direct precipitation in open water during intervals of past ice-sheet collapse during Marine Isotope Stage 5e (MIS-5e; 120 kyr BP) or earlier Pleistocene interglacials³⁵. These microfossils predate any measurable radiocarbon source in the sub-glacial environment.

Bulk sediment samples for radiocarbon measurements were wet sieved with nanopure water through a 63-\$\mu m\$ screen to remove coarser mineral matter, then pre-treated using standard acid-base—acid protocols\$^36\$. The remaining insoluble fraction for each sample was combusted to convert to CO2 and then graphitized. Samples were then measured using accelerator mass spectrometry at the WM Keck Carbon Cycle Laboratory at the University of California, Irvine. A subset of samples was also independently pre-treated and measured at the Uppsala radiocarbon facility, Sweden, following the same protocol. Owing to the inherent age uncertainties, radiocarbon 'ages' are presented as raw, uncalibrated values that are not corrected for known reservoir effects. Acid-insoluble organic 13 C ratios were generated separately at the Environmental Isotope Laboratory at the University of Arizona, following standard methods.

In order to minimize the potential for contamination of small samples during analysis, we processed large samples (larger than 150 mg), producing around 1 mg to 2 mg of carbon that was combusted and reduced to graphite, while simultaneously processing numerous primary and secondary standards, ranging in size from very small (less than 0.1 mg) to large (1.8 mg). To further demonstrate that we have thoroughly explored a wide range of radiocarbon systematics, we also dated base-soluble fractions for a subset of samples. The base-soluble fraction (humic acid) resulted in ages that were somewhat older (about 1–2 kyr old) than the bulk sediment samples. These age differences are relatively small and the results are consistent with our other findings. Multiple fractions that yield similar ages further rule out contamination as a possible explanation of our radiocarbon results.

We considered and ruled out sample contamination by modern carbon before analysis as a potential explanation of the radiocarbon results. Subglacial samples were recovered between 1989 and 2013, and sub-ice-shelf samples were recovered in 2015 (WGZ) and 1978 (RISP). SLW (2013) samples were recovered and handled using full clean-access protocols³⁷, where all instruments were peroxide washed before recovery. Cores and samples were sealed and maintained in a +4°C environment. Clean-access protocols were not used during earlier sample recovery (for WIS, KIS and BIS), although every effort was made to maintain appropriate cleanliness in the field and in the laboratory. These samples were sealed and maintained at +4 °C. Subsamples were stored in sealed sections of plastic-core liner, in plastic bags or in plastic vials. Many of the samples in vials dried out and some of the dried samples have been stored at room temperature in the intervening years. The fact that the older subglacial samples demonstrate the oldest apparent radiocarbon ages argues against any introduction of new carbon from microbial or fungal growth on the sample. RISP cores were stored at the Florida State University Antarctic sediment-core repository, sealed and chilled. The somewhat younger ages there are readily explained by the long-term exposure to the sub-ice-shelf ocean cavity. Despite the different sample storage methods used, the radiocarbon results are very consistent, which argues against contamination. Given the small concentration of organic carbon in the samples, a very small amount of contamination with modern carbon would result in some anomalous younger ages, yet all of our results fall within a narrow range. Given the range of sample sources and storage methods, equal contamination of all the samples consistent with our results is extremely unlikely.

Apparent and true ages of sediments. Given the dominant concentration of old (radiocarbon dead) organic carbon in the samples³⁸, all ages presented here are older than the likely age of the pure radiocarbon component; note the calculated 'per cent modern' column in Extended Data Table 1. We infer post-LGM ages for all samples. Apparent radiocarbon ages for 11 till samples from beneath the Whillans, Kamb and Bindschadler ice streams were obtained for acid-insoluble organics (AIOs) and span from around 20 to 35 kyr BP (Extended Data Fig. 5).

Rare, small biogenic carbonate fragments from molluscs, foraminifera and calcareous nanofossils have been found in several samples, but these are all Tertiary in age (on the basis of biostratigraphic assessment¹⁹), and we made no attempt to radiocarbon date them.

Sediments recovered from beneath the southern Ross Ice Shelf (RISP cores; Fig. 1) were radiocarbon dated, generally yielding somewhat younger ages than the ice streams, probably reflecting the longer period of contact with the sub-ice-shelf marine cavity. For the most part, the raw ages appear to correspond to the LGM of the WAIS, which started about 29 kyr BP and ended 13.9–15.2 kyr BP (ref. ³⁹). Many lines of geologic evidence document that the LGM grounding line of the WAIS was located at or near the Ross Sea continental shelf break⁴⁰ at the time that corresponds with the apparent, uncorrected radiocarbon ages in our samples.

For the apparent ¹⁴C ages to represent the true sediment ages, the grounding line of the WAIS would have to be upstream of the core sampling locations and also require all of the carbon pool in the samples to initially have had the standard, modern 14C/12C ratio. However, owing to the large oceanic reservoir effect in Antarctica, even modern amphipods sampled by us in January 2015 through a borehole at the grounding line of Whillans Ice Stream (WGZ; Fig. 1 and Extended Data Table 1) had a fraction of modern radiocarbon of only 0.8669-0.8746, corresponding to apparent ages of 1,075-1,145 ¹⁴C years. Moreover, it is well documented that radiocarbon dates on AIOs obtained from bulk Antarctic glacigenic sediments are typically biased by admixture of old, ¹⁴C-depleted organic matter^{4,41–43}. This old organic material comes from glacial erosion of sediments deposited earlier in the $Cenozoic^{34}$. The tills of the Ross ice streams are dominated by Tertiary, mostly Upper Miocene^{34,35}, marine source beds that are being actively eroded by grounded ice¹⁹. Given the uncertainty concerning the initial mixture between 'young' and 'old' sources of organic matter, we know only that the true age of the radiocarbon falls somewhere along the exponential-decay lines of ¹⁴C in Extended Data Fig. 5, which intersect the left-hand vertical axes of this figure at the measured values of the ¹⁴C modern fraction.

Instead of the apparent ¹⁴C ages representing the true sediment ages, it is more reasonable to assume that the WAIS grounding line was upstream of the till sampling locations after the LGM, and that the calculated ages are biased towards older dates owing to the high concentration of ancient carbon (Extended Data Fig. 5). Our assumption is also compatible with the relatively low initial fractions of ¹⁴C in sampled sediments, which we expect given that the sampled subglacial areas were exposed to an influx of marine-sourced radiocarbon over a geologically short period of time and were located very far from the main locus of regional biological productivity in the Ross Sea. For instance, if our sediment samples had received ¹⁴C-bearing marine organics in the mid-Holocene (around 5 kyr вр), then the initial fractions of ¹⁴C for these samples could be quite low (roughly 0.03 to 0.14) to explain the obtained apparent ages. By contrast, if one chooses a time period pre-dating the WAIS LGM, say 30 kyr BP, then most of our samples would need to have all of their organic matter completely equilibrated with the oceanic pool of ¹⁴C at that time. Such conditions are difficult to find even in the modern open marine sediments of Ross $Sea^{41,42}$.

The balance of evidence favours a post-LGM origin of ¹⁴C-bearing organics in our till samples. However, the radiocarbon data do not allow us to pinpoint more precisely when the proposed retreat and re-advance of the WAIS grounding line took place in the Ross Sea sector of the ice sheet, or the specific duration of exposure.

Potential input of radiocarbon from basal melting. Here we check whether 14C in subglacial sediment samples from beneath three different ice streams may have been entirely, or at least to a substantial extent, supplied by basal melting of meteoric ice. Meteoric ice may contain as much as 140 mm³ of air per gram⁴⁴, which translates into about 0.02 g of total inorganic carbon (TIC) per m³ of ice, assuming an ice density of 910 kg m⁻³ and a pre-industrial Holocene atmospheric concentration of carbon dioxide (280 parts per million, p.p.m.). Meteoric ice also contains organic matter deposited from the atmosphere 45,46. From these two publications $^{45,\bar{4}6}$ we select 100 μg per litre as an upper bound on the total organic carbon (TOC) concentration in ice coming from the interior of the ice sheet. This assumption yields around 0.1 g of TOC per m³ of ice. TIC and TOC combined give 0.12 g of carbon per m³ of meteoric glacial ice. Basal melting rates vary beneath the Ross Sea sector of the WAIS, but 0.003 m per year provides a representative estimate for the region⁴⁷. At this rate 0.36 g of carbon per m² of the bed area would be entering the subglacial zone of the ice sheet in each thousand years. Some of this material would be entering subglacial sediments already as organic carbon melted out of the ice, whereas the component derived from carbon dioxide trapped in the melting ice would exist as dissolved inorganic carbon (DIC). We assume that the latter could be relatively quickly sequestered by subglacial microbial activity and converted into organic matter48.

The basal flux of carbon estimated above needs to be compared with the total stock of carbon in the subglacial till from which our samples are derived. Our radiocarbon measurements show that ¹⁴C is present at least within the top metre of till

recovered from beneath three Ross Sea sector ice streams. Analyses performed in the University of California Santa Cruz stable-isotope laboratory (UCSC CF-IRMS) on 27 subglacial sediment samples show an average TOC of 0.33% (with a standard deviation of 0.14%, both expressed in weight per cent of the dry sedimentary matter). Because the dry density of the till is about 1,600 kg m $^{-3}$ (ref. 49), a metre-thick layer of till contains about 5 kg of organic carbon per m 3 of sediment. Even if we assume that all of the carbon entering the subglacial zone with basal meltwater is sequestered within the top metre of till, it would take about 14 million years to supply the total amount of carbon found in these sediments just from basal melting at the rate of 0.36 g per 1,000 years. Owing to $^{14}\mathrm{C}$ decay, only the carbon released by basal melting during the last tens of thousands of years can contribute to the present stock of this radioisotope in till. The remaining ratio, *R*, of undecayed $^{14}\mathrm{C}$ in a pool of carbon accumulating through time, *t*, by the addition of new $^{14}\mathrm{C}$ -bearing matter at a constant rate can be calculated from:

$$R_{\{t\}} = \frac{R_0}{t} \int_0^t e^{-\varsigma/\lambda} d\varsigma = R_0 \frac{\lambda}{t} (1 - e^{-t/\lambda})$$

where R_0 is the initial 14 C ratio (for example, the modern atmospheric 14 C 12 C ratio), ς is a dummy variable of integration, and $\lambda = 3,972$ years (a constant given as the product of 14 C half-life, 5,730 years, and the natural logarithm of 2). After 14 million years (Myr), the hypothetical subglacial carbon pool resulting solely from a continuous accumulation of carbon released from basal melting of meteoric ice would have an average 14 C ratio of only 0.00028 of its initial (for example, modern) value. This is two orders of magnitude too low to explain the fractions of 14 C measured in our samples. From the equation above, we can calculate that the observed fractions of 14 C could be explained only by constant accumulation of carbon with modern initial 14 C over periods of time of around 100 kyr or less. However, the flux rate of carbon from basal melting would then have to be around 50 g per thousand years per unit area of the ice base in order to explain the total stock of carbon in the sampled subglacial sediment layer (around 5 kg m $^{-3}$). As per our discussion above, such rates of carbon delivery from melting basal ice are implausibly high.

The analyses presented here did not even take into account the fact that any carbon released from the base of the ice streams has spent thousands to tens of thousands of years stored in the ice itself, which would further decrease its ¹⁴C content. Furthermore, an ice-stream base can also be composed of basal ice that has been formed by the freezing of subglacial waters. Such basal ice would not contain ¹⁴C-bearing carbon dioxide or organic matter. Hence, we conclude that the release of ¹⁴C from the base of the ice sheet does not represent a substantial source, and that the inclusion of recent marine organic matter during a recent ice sheet retreat is needed to explain the concentrations of this isotope measured in our samples. Furthermore, the radiocarbon results we report are completely consistent with the ice-sheet retreat ages inferred from the radar profiles and modelling reported here. A modern analogy from the present-day sub-shelf cavity. Hot-water drilling through 760 m of ice into 10 m of water in a sub-ice shelf-embayment more than 600 km from the open ocean, at the Whillans Ice Stream grounding zone⁵⁰ (WGZ; Fig. 1), revealed a diverse community of organisms—including diverse amphipods, zoarcid and notothenioid fishes, and medusoid and ctenophorid jellies—thriving in fully marine water. Radiocarbon analysis of appendages from three live-captured amphipods yielded raw ages between 1,075 \pm 20 and 1,145 \pm 20 yr BP (Extended Data Table 1), comparable to the Ross Sea surface water reservoir age⁵¹. The results from this grounding-line-proximal community of organisms show that radiocarbon is introduced from the open ocean virtually everywhere that ocean waters reach beneath the ice shelf. A retreating grounding line would have opened a subglacial marine environment that was immediately colonized by organisms that leave a radiocarbon tracer on their death. This modern sub-ice shelf process illustrates a likely pathway for Holocene radiocarbon to be deposited upstream of the present grounding line following past grounding-line retreat. Furthermore, pore-water chemistry indicative of seawater at Subglacial Lake Whillans (SLW; Fig. $1)^{48}$ demonstrates that marine waters previously occupied the subglacial lake basin.

Henry Ice Rise: observations, interpretation and flow history. Henry Ice Rise (HIR) is one of several ice rises in the Weddell Sea that influence the flow of the Ronne Ice Shelf and its ice streams. It is currently slow flowing³⁰ and cold based⁵². On the basis primarily of new ground-based ice-penetrating radar data, we hypothesize that HIR formed during the Holocene as the Ronne Ice Shelf grounded on a bathymetric high. Here we describe the radar system and our processing steps, and discuss possible links between surface roughness and englacial structure, which pertain to a potential past grounding-line configuration. We also discuss an alternative interpretation that HIR existed throughout the Holocene, but was in the past smaller than it is today.

Radar system. We used the British Antarctic Survey's DEep LOoking Radio-Echo Sounder (DELORES) on HIR to map basal topography and englacial structure⁵³. A transmitter producing 2,500 broadband radiowave pulses per second was

connecting to a 20-m, resistively loaded dipole antenna, so that the centre frequency of the system in ice was 4 MHz. A receiver unit, positioned 100 m from the transmitter and connected to an identical dipole antenna, was triggered by the air wave and sampled the return signal at 250 MHz. The system was towed 50 m behind a snowmobile, driven at about 15 km $\rm h^{-1}$. After stacking, this configuration produced traces roughly every 85 cm along the track.

Data processing. Traces were geolocated in three dimensions using data from a dual-band GPS unit, mounted at the midpoint of the transmitter and receiver, then interpolated onto a regularly spaced grid, bandpass filtered and compiled into radargrams. Radargrams were migrated with a two-dimensional Kirchoff scheme, assuming a constant radiowave velocity of $0.168 \, \mathrm{m} \, \mathrm{ns}^{-1}$ (ref. ⁵³).

Elevations of the ice-bed interface and one of many englacial reflecting horizons, interpreted as isochrones, were determined using the software package Petrel by Schlumberger. Conversion from the two-way travel time of the radar signal to depth was made assuming a constant radiowave velocity (0.168 m ns $^{-1}$). Correcting for the impact of the lower density of the firn on radiowave velocity would decrease the elevation by up to 10 m. As firn densities are unknown on HIR and probably vary spatially, we plot the uncorrected elevation of the ice-bed interface, $z_{\rm b}$ (Fig. 2a). Also plotted in Fig. 2a are ice-bed elevations from the BEDMAP2 dataset (H. Pritchard, personal communication). The normalized elevation of the isochrones, $\zeta_{\rm i}$, is computed from $\zeta_{\rm i}=(z_{\rm i}-z_{\rm b})/(z_{\rm s}-z_{\rm b})$, where $z_{\rm i}$ is the elevation of the isochrone and $z_{\rm s}$ is the ice-surface elevation measured with the dual-band GPS (Fig. 2c).

Surface-texture contrasts visible in satellite imagery and surface-elevation data. Antarctic satellite imagery can be used to reveal subtle ice-surface topography^{54,55}. The MODIS Mosaic of Antarctica image in Fig. 2 highlights contrasting regions near the northern end of HIR, which we interpret as indicative of contrasting surface roughness. The green curves in Fig. 2b, c highlight the boundaries between the regions. The area between the two green curves appears smoother than the two regions between the green curves and the present-day grounding line. This interpretation is consistent with surface slopes estimated from elevation data collected by the GPS unit mounted on the DELORES radar system (data not shown).

The surface-texture contrasts are approximately parallel to the present-day grounding line and align roughly with the following features revealed by our ice-penetrating radar survey: extensive synclines in internal isochrones (Fig. 2c and Extended Data Fig. 2a), locations where isochrones intercept the bed (Fig. 2b and Extended Data Fig. 2f), and the onset of buried crevasses (Fig. 2b). Here we explain these alignments by proposing that all of these features are the signatures of a past grounding-line configuration that persisted during a period either following the grounding of the ice rise, or when HIR was at its minimum extent (see below).

Today on the ice-shelf side of the eastern grounding line of HIR, ice undergoes lateral shear as the ice shelf moves past the relatively slow ice rise. This shear generates a region of dense crevassing (Extended Data Fig. 2g). Deformation in shear margins also warms englacial ice and generates ice-crystal fabric. Both can influence the effective viscosity of ice, as can crevassing. As the grounding line swept through the region between the surface-texture contrast and the present-day grounding line, crevasses generated on the ice-shelf side would have become inactive, buried and then deformed in the slow-moving ice. Simultaneously, englacial temperatures and crystal fabrics would have evolved in a complex manner as the shear margin migrated in step with grounding-line migration. We hypothesize that these changes, together with spatially heterogeneous basal melting, resulted in the complex pattern of tilted crevasses we observe today.

Under this interpretation, the rougher surface texture in the region presently occupied by buried crevasses (Fig. 2) results from spatially variable ice viscosity caused by variability in the orientation and height of crevasses, as well as spatially variable ice fabric and temperature that have evolved enough to still affect ice flow. By contrast, the ice in the region between the two green curves (Fig. 2) has undergone a simpler flow history, without substantial lateral shearing, either because it was immediately upstream of the initial location of grounding (the subglacial high in Fig. 2a) and experienced only longitudinal compression, or because it did not unground. We discuss the latter scenario next.

An alternative interpretation: a smaller-than-present but persistent HIR. In the main text and in the previous section, we interpreted our radar observations to indicate that HIR became completely ungrounded following the LGM, then formed through regrounding on the topographic high at the northern end of HIR. However, some of our radar observations can be explained by an alternative ice-flow history. The grounding line surrounding HIR may have retreated substantially, exposing areas of the ice base to the ocean, where heterogeneous basal melting deformed and truncated isochrones at the bed. Subsequent re-advance of the grounding line would have buried crevassing as described above. This would have had an effect on regional ice-shelf dynamics and buttressing, but HIR would have persisted throughout the Holocene. However, if the locations where isochrones are truncated at the bed correspond to areas that ungrounded, then the simplest minimum extent suggested by mapping the layer truncations (Fig. 2b) would involve the ice

rise ungrounding over the highest basal topography (Fig. 2a), while remaining grounded over deeper bathymetry. We do not yet understand ice-rise dynamics sufficiently to fully assess if this is possible. However, such a pattern of ungrounding during deglaciation is inconsistent with recent numerical modelling of idealized ice-rise formation⁵⁶. Therefore, we argue that full ungrounding and later regrounding is more likely than partial ungrounding.

Whether the ice rise ungrounded completely or partially, the buttressing force exerted by the ice shelf on the ice sheet upstream would have been affected. These scenarios could be tested by drilling to the ice-rise base to obtain sediments for radiocarbon analysis and to allow measurement of the englacial temperature profile⁵⁷.

Ice-sheet modelling. *Model description and forcings.* We used the open-source Parallel Ice Sheet Model (PISM)^{23,58,59} to perform pan-Antarctic simulations with glacial-cycle climate forcings. PISM is a three-dimensional, thermo-mechanically coupled ice-flow model with a freely evolving grounding line and calving front. The hybrid shallow approximation of Stokes flow allows for large-scale, longterm simulations of ice-sheet evolution. Unless otherwise stated, we used surfacetemperature anomalies from the WAIS divide ice-core (WDC) reconstruction⁶⁰, which show a sharp increase of 11 K starting around 17 kyr BP. For surface accumulation, we use the 1980-2000 mean accumulation from the output of a regional climate model (RACMOv2.1, HadCM3; ref. 61) as a base accumulation pattern and scale this pattern by 2% per degree of climatic temperature change from present⁶² (using the WDC reconstruction) and by 43% per km of surface elevation change. The latter assumes a linear dependence of air temperature on elevation combined with an exponential dependence of precipitation on temperature. At the ice-ocean interface we use the Potsdam Ice-shelf Cavity mOdel (PICO)⁶³, which calculates melt patterns underneath the ice shelves for given ocean conditions⁶⁴. Ocean temperature anomalies are computed from ice-core derived surface-temperature anomalies convolved with a response function to produce a damped and delayed response⁶⁵. The calving front can freely evolve with calving parameterized to be dependent on principal strain rates at the ice-shelf front⁶⁶. Basal sliding is parameterized using an iterative optimization scheme⁶⁷ modified for the till-friction angle, mimicking the distribution of marine sediment and bedrock, such that the mismatch to modern surface elevation observations is minimized.

Sea-level change drives grounding-line migration through the flotation criterion, which determines grounding-line position ⁶⁸. We prescribe sea-level changes by considering the height of the sea surface and the height of the sea floor separately. Unless otherwise stated, we use global mean sea-surface heights prescribed by the ICE-6G GIA model ⁶⁹. According to this model, mean sea-surface height has risen by about 100 m since 14.5 kyr BP. Alternative sea-surface height records were considered as part of the sensitivity analysis discussed below.

Changes to the height of the sea floor and bed topography are modelled using an approach that reflects the deformation of an elastic plate overlying a viscous half-space. Calculations are carried out using the computationally efficient Fast Fourier Transform to solve the biharmonic differential equation for vertical displacement in response to ice-load change⁷⁰. This approach can also be used to calculate vertical displacement in response to spatially varying water-load changes (more details below). A key advantage this approach has over traditional Elastic Lithosphere Relaxing Asthenosphere (ELRA) models is that the response time of the sea floor is not considered a constant, but depends on the wavelength of the ice-load perturbation. This formulation closely approximates the approach used within many GIA models⁷⁰. Given that our ice-sheet model is not coupled to a GIA model, we are unable to prescribe self-consistent water-load changes or account for feedbacks associated with post-glacial changes to the rotational state of the Earth⁷¹. The effect of neglecting these processes is discussed below in the section on sea-level forcing.

Bed-elevation adjustment. With a resolution of 15 km and uncertain bed elevation, basal conditions and climate forcings, matching the present-day grounding-line position in the Weddell Sea required raising the ice-sheet bed in one key location (Bungenstock Ice Rise) to compensate for topographic information lost during remapping.

The present-day elevations of the sea bed and ice-sheet bed are regionally highly uncertain ^{29,72}. Furthermore, when remapping observed bed elevations (Bedmap2; ref. ²⁹) from a relatively fine spatial grid (1 km) to the spatial resolution of our simulations (15 km), we lose bed-elevation information in key places. Remapping introduces inherent uncertainty into any low-resolution ice-sheet modelling study, but it is particularly important for the process of ice-rise regrounding that we highlight. For example, at present-day ice rises the remapping of the bed-elevation data reduces the apparent peak bed elevation by 36–135 m, while at their steep flanks this difference can be a few hundred metres (Extended Data Fig. 4). We find that in our simulations, if we use bed topography remapped directly from the Bedmap2 compilation (using a first-order conservative technique⁷³), the grounding line in the Weddell Sea sector does not re-advance across a 1,300m-deep trough and

often remains near to its Holocene minimum position, far inland of its present-day location, until the end of simulations. This is unrealistic.

We have experimented with various approaches to dealing with the uncertainty introduced by remapping bed topography to lower spatial resolutions. These include adopting the maximum Bedmap2 value in each model grid cell, either in the regions of individual ice rises or across the whole ice sheet. We also experimented with a sub-grid pinning point scheme, dependent upon the thickness of the water column underneath the ice shelf within some uncertainty range⁷⁴ and with a simpler uniform adjustment in the region of individual ice rises. Which approach we take affects the timing and magnitude of grounding-line retreat and re-advance. Without a clear motivation to adopt a more complex approach, we made the minimum adjustment to the bed that allowed the grounding line to re-advance in the Weddell Sea sector: we uniformly raised the bed by 150 m in a 165 km by 180 km area centred on Bungenstock Ice Rise (BIR) only. This rather arbitrary choice is a major limitation of this model ensemble, which (along with other uncertainties associated with model resolution, forcings, parameters and physics; see below) prevents us from extracting information about the timing of grounding-line retreat and re-advance from our simulations.

The purpose of our model experiments is to explore the mechanisms that could have caused re-advance and what influences these mechanisms. It is beyond our scope to explore the range of options to compensate for basal topographic remapping errors, but our work highlights that, at least for studying ice-rise regrounding, resolving this issue will be required if we are to make quantitative predictions of millennial-scale ice-sheet behaviour.

Model ensemble and the reference simulation. We performed an ensemble of simulations, each spanning 205 kyr BP to the present, in which uncertain parameters were systematically varied and the results were compared with palaeo-ice-sheet datasets and present-day observations^{75,76}. The full results of the ensemble represent a likely range of Antarctic ice-sheet chronologies and will be presented elsewhere. Here we focus on the possible extent and triggers of large-scale grounding-line re-advance during the Holocene, and so discuss in detail only those mechanisms that are relevant to this process. We choose one of the ensemble members to act as a reference simulation to demonstrate aspects of model behaviour. The reference simulation is chosen from many ensemble members that use parameters lying within physically plausible bounds (Extended Data Table 2) and which also achieve reasonable agreement with a commonly used ice-sheet grounding-line position reconstruction². Although this grounding-line reconstruction does not include grounding-line re-advance during the Holocene, as discussed in the main text, many ensemble members—including our reference simulation—simulate the grounding line retreating substantially inland of its present-day position and subsequently re-advancing towards its present position. We express ice-mass changes as the above-flotation volume in units of global sea-level equivalent, assuming a constant ocean area of 3.61×10^{14} m² (ref. ⁷⁷).

The drivers of grounding-line re-advance. We performed three model experiments (separate from the full ensemble, above) in order to disentangle the causes of re-advance in the Weddell and Ross Sea sectors. We find that both uplift of the bed at the grounding line and buttressing caused by the formation of ice rises drive re-advance of the grounding line towards its present-day position in both the sectors (Extended Data Fig. 4). The first experiment ('No uplift'; Extended Data Fig. 4) is identical to the reference simulation except that uplift is halted after 10 kyr BP—that is, at approximately the time at which the grounding line in the reference simulation reaches its most retreated position in both sectors (Fig. 3). The grounding line in the Weddell Sea remains at its 10 kyr BP position for the remainder of the simulation. In the Ross Sea the grounding line retreats further into the interior of the ice sheet. This additional retreat can be prevented by buttressing, as demonstrated in the second experiment ('No uplift, grounding of ice rises'; Extended Data Fig. 4), where uplift is again halted at 10 kyr BP, but ice-rise formation is enforced by raising the seafloor in the locations of the Crary, Steershead, Henry and Korff ice rises and Doake Ice Rumples. In this simulation, further retreat in the Ross Sea is prevented, but re-advance still does not occur in either sector. We further test the relevance of buttressing via ice-rise formation in a third experiment ('Uplift, no grounding of ice rises'; Extended Data Fig. 4), in which uplift of the bed is allowed, but ice-rise formation is prevented by lowering the seafloor. In this simulation ice-shelf buttressing is reduced compared with the reference simulation. Consequently, the grounding line remains at its 10 kyr BP position in the Weddell Sea (Extended Data Fig. 4a) and relatively little re-advance occurs in the Ross Sea (Extended Data Fig. 4b). Hence we identify the grounding of HIR, as evident from our radar survey, as critical for grounding-line re-advance in the Weddell Sea in these simulations; meanwhile, in the Ross Sea, neither uplift in the grounding-line region nor buttressing resulting from ice-rise formation is alone sufficient to drive grounding-line re-advance to the present-day position. Model sensitivity to forcings. Extended Data Fig. 6 plots selected results from our analysis of the sensitivity of the model to various forcings. The retreat of the grounding line inland of its present-day location and its subsequent re-advance

is a common behaviour in the model; however, its minimum extent during the Holocene, and how fast and how far it re-advances, are all sensitive to forcings.

Given that sea-level forcing is highly relevant for deglaciation, we compared the responses to four different eustatic sea-level reconstructions (Extended Data Fig. 6a). In our reference simulation, we use the sea-level curve from the ICE-6G model⁶⁹. Refs ^{78,79} provide similar sea-level reconstructions and hence similar model results, with the strongest changes after around 15 kyr Bp. We also used the SPECMAP time series⁸⁰ (as used in the SeaRISE intercomparison⁷⁷); the results show a delayed LGM sea-level lowstand and a delayed sea-level rise to Holocene conditions, and hence the modelled ice sheet exhibits a later retreat and re-advance (Extended Data Fig. 6).

In order to mimic the first-order effects of GIA coupling⁸¹ (including rotational feedback and self-gravitational effects), we experimented with scaling the time series of sea-level forcing by factors of 0.9 and 0.8 (initiated at 35 kyr BP; Extended Data Fig. 6b). We do not attempt to prescribe spatially varying sealevel forcing, but comparison with independent GIA model output 82 suggests that neglect of rotational feedback and self-gravitation of the ocean may result in local errors in sea-level forcing of the order of 15-20 m (this range reflects the likely error associated with prescribing sea-surface height; deformation of the seabed is self-consistently modelled within PISM). Scaling the uniform sea-level forcing by a factor of 0.9 causes the lowstand to be less pronounced at the LGM (approximately 10 m higher), in comparison with the reference simulation. This affects the LGM grounding-line position, particularly in the Ross Sea, which in turn affects the retreat and re-advance of the grounding line, because the depression of the bed depends sensitively on the ice-sheet's LGM extent. Scaling by 0.8 may be unrealistic (on the basis of comparison with GIA model output generated using an independent ice-sheet history⁸²; data not shown), and interestingly, we note that retreat behind the present-day grounding-line position is not reproduced in this scenario (Extended Data Fig. 6b). We also experiment with a sea-level forcing that is identical to that used in the reference simulation (ICE-6G model⁶⁹) except that the curve has been uniformly shifted 2 kyr earlier. The result is that the grounding line responds with an earlier retreat. This response emphasizes the key role of the sea-level forcing in triggering large-scale grounding-line retreat.

Sea-level changes also affect the load of the ocean on the sea bed. This triggers bed deformation, which will affect grounding-line migration. By default, this second-order effect is not accounted for in PISM. However, we carried out exploratory simulations that do account for it (data not shown), and find that when grounding-line retreat is accompanied by an increase in eustatic sea level, the additional ocean load partly counteracts the unloading associated with grounding-line retreat. Accordingly, the grounding line retreats further inland than in the reference simulation. On the other hand, sea-bed uplift following grounding-line retreat reduces the water load in marine sectors; this further amplifies uplift, which supports grounding-line advance. These interesting second-order effects do not qualitatively affect model behaviour, but they do influence the magnitude of grounding-line retreat and re-advance through their influence on LGM extent in both the Weddell and Ross sectors.

For surface-temperature forcing, our reference simulation uses a reconstruction based on data from the WAIS divide ice core (WDC) 60 . The results are similar when an alternative reconstruction from the EPICA Dome C ice core (EDC) 83 is used (Extended Data Fig. 6c). However, the grounding line responds to the slightly warmer LGM conditions in the EDC case with less LGM advance and hence a less severe retreat in the Ross Sea sector. For comparison, we also force one simulation with a temperature record pertaining to the start of the Last Interglacial Period (from the EDC core), in which an earlier and stronger warming leads to an earlier and stronger grounding-line retreat, particularly in the Ross Sea sector.

Accumulation in the reference simulation is coupled to changes in surface temperature by imposing a 2% precipitation change for each degree of variation from present-day temperatures (Extended Data Fig. 6d; violet curves) resulting from climatic changes (constrained by ice-core data), and a 43% precipitation change per km of surface-elevation change. We experimented with two alternative timedependent accumulation forcings and two constant accumulation scenarios. Using either a scaling of 5% per degree of WDC-temperature change or an independent WDC-derived accumulation reconstruction⁸⁴ leads to lower mean accumulation and a less advanced LGM grounding-line position (Extended Data Fig. 6d). The less advanced LGM grounding line almost eliminates grounding-line retreat inland of its present position and re-advance, particularly in the Weddell Sea. When accumulation is kept constant at LGM conditions (2% per degree scaling of the EDC temperature at 25 kyr BP; Extended Data Fig. 6d), which correspond to 18% lower accumulation than today, the grounding line retreats inland of its present-day location in both sectors, but only partially re-advances in the Ross Sea and does not re-advance in the Weddell Sea. When accumulation is kept constant at present-day values (Extended Data Fig. 6d), grounding-line retreat starts earlier than in the reference simulation, particularly in the Ross Sea. In both sectors the grounding line retreats and re-advances in a similar way to the reference simulation, but with different timings: in the Weddell Sea the grounding line re-advances several thousand years earlier than in the reference simulations, while in the Ross Sea re-advance is delayed in comparison with the reference simulation.

Model sensitivity to parameters. Next we used selected members of the ice-sheet model ensemble to demonstrate the sensitivity of the model to various parameter values. Analysis of the full model ensemble, including a systematic validation of the full range of parameter combinations against present-day conditions and reconstructions of past conditions^{27,76}, will be presented elsewhere. Here we present the impact of single parameter perturbations. In general, we find that retreat of the grounding line inland of its present-day location and subsequent re-advance occurs over a wide range of parameter choices, but the Holocene minimum extent, and how fast and how far the grounding line re-advances, are sensitive to these choices.

Mantle viscosity affects model behaviour because it defines the rate and pattern of the deformation of the ice-sheet bed and sea floor. Our reference simulation uses a mantle viscosity of 5×10^{20} Pa s. The ensemble also covers a value considered typical for pan-Antarctic model simulations and often used as the default value in other PISM simulations (1×10^{21} Pa s; Extended Data Fig. 7a). We selected the lower value for our reference simulation to account for the weaker mantle beneath the WAIS 85 . An even lower viscosity (roughly 1×10^{20} Pa s; Extended Data Fig. 7a) has also been tested. In the lowest viscosity case, retreat of the grounding line inland of the present-day position is prevented as the bed responds too quickly to ice unloading. We find the fastest grounding-line retreat rates for higher viscosities. The most inland position reached by the grounding line is similar in each case, except the lowest viscosity case, and re-advance occurs earlier for lower mantle viscosity. In summary, we find that grounding-line retreat and re-advance occurs in a plausible but confined range of mantle viscosity values.

Flexural rigidity is associated with the thickness of the elastic lithosphere and has an influence on the horizontal extent to which bed deformation responds to changes in load. Previous studies based on gravity modelling suggest appropriate values for our study with a focus on West Antarctica lying within the range 5×10^{23} to 5×10^{24} N m (refs 86,87). Our reference simulation marks the upper end of this range (Extended Data Fig. 7a). For lower values, 1×10^{24} to 5×10^{23} N m, we find grounding-line retreat beyond its present-day location and re-advance as in the reference simulation. However, maximum retreat is delayed in the Ross Sea sector, so re-advance of the grounding line does not reach its present-day location in that sector.

Enhancement factors are used in ice modelling to account for anisotropy and other unresolved rheological properties and enter the constitutive law. PISM uses one enhancement factor for the shallow-shelf approximation (SSA) component of the stress balance, and a second enhancement factor for the shallow-ice approximation (SIA) component. Increasing the SSA enhancement factor (Extended Data Fig. 7b) and/or decreasing the SIA enhancement factor (Extended Data Fig. 7c) produces a less advanced LGM grounding-line position. This is because larger values of the SSA enhancement factor produce faster ice streams and thinner ice shelves, and smaller values of the SIA enhancement factor produce thicker grounded ice. For a less advanced LGM grounding line, retreat begins earlier and progresses more slowly, and does not reach as far inland before retreat is halted.

PISM uses a generalized sliding parameterization formulated as a power law⁵⁹ spanning a range from plastic Coulomb sliding (with sliding exponent q=0) to sliding in which till strength is linearly related to sliding velocity (q=1). In the reference simulation we use q=0.75 (Extended Data Fig. 7d). In the linear case (q=1), the LGM grounding line is less advanced and retreat starts earlier (Extended Data Fig. 7d). For smaller values of q, retreat occurs generally later in the Weddell Sea and retreat in the Ross Sea is less pronounced.

Two other parameters associated with the sliding parameterization are the decay rate of till water and the effective overburden pressure⁵⁹. Within the range explored by the ensemble, both parameters have only a moderate effect on the LGM extent of the grounding line and the timing of retreat, and do not affect whether or not the grounding line retreats inland of its present-day location and re-advances (Extended Data Fig. 7e).

A final sliding-related parameter is the till friction angle, which varies spatially and for our reference simulation is optimized 67 to minimize the mismatch between modelled and observed surface elevation, but is constrained to be larger than 2°. Reducing the minimum value to 1° leads to a smaller LGM extent and hence a slower retreat and larger minimum extent (preventing retreat past the present-day grounding-line position in the Weddell Sea) (Extended Data Fig. 7f). Instead of optimizing the till friction angle using observed surface elevations, it can also be defined as a linear piecewise function of bed topography, with 2° used in areas below $-500\,\mathrm{m}$ (this is the default approach in PISM) 59 . This also reduces the LGM extent and, in the Ross Sea, reduces the retreat of the grounding line inland of its present-day extent.

Ocean forcing in our simulations is modelled with PICO⁶³. PICO uses parameters for overturning strength and heat exchange. Modification of the parameter values affects the LGM extent of the grounding line and hence the rate and timing

of retreat (Extended Data Fig. 7g). However, grounding-line retreat and re-advance are produced as robust features for extreme parameter values, even if melting is omitted or prescribed as a constant at present-day values.

Calving is parameterized as eigencalving (dependent on strain rates)⁶⁶. A parameter K is the constant of proportionality between the calving rate and the horizontal spreading rate of ice shelves (Extended Data Fig. 7h). K is assumed to be constant and uniform. Our reference simulation uses $K=1\times 10^{17}$ m s. The LGM grounding-line position is less advanced for smaller eigencalving values, and grounding-line retreat less pronounced, probably because of additional ice-shelf buttressing resulting from less calving.

Resolution dependence. Our simulations, in common with all millennial-timescale ice-sheet simulations, suffer from major limitations related to the maximum practical spatial resolution that they can use. Just like the model parameters considered in the previous section, the spatial resolution can be treated as a quantity that affects the results of the simulations and should be investigated. This is particularly true in our study, as ice-shelf grounding on bathymetric highs with relatively small horizontal dimensions has proven to be so important for the large-scale evolution of the ice sheet.

A sensitivity analysis aimed at examining the sensitivity of this behaviour to resolution (analogous to the exercise described above) is highly limited by computational resources. For example, doubling the spatial resolution incurs at least a tenfold increase in computational cost. Ensembles with systematically varied parameters of simulations that span the full spin-up over two glacial cycles (205 kyr) are at present possible only with a spatial resolution of 15 km.

Shorter simulations (that cover only the past 20 kyr) are possible using a resolution of up to 7 km, if they are initiated at 20 kyr BP by remapping the spun-up state of a 15-km-resolution simulation. (Unfortunately, this remapping means that, despite the higher resolution, the bed topography is no better resolved with respect to observations²⁹ than the 15-km-resolution simulations.) Higher-resolution simulations generally reproduce the pattern of grounding-line retreat and readvance, but the increase in resolution strongly affects the timing and magnitude of changes (Extended Data Fig. 8). Owing to the influence of resolution on other model parameters, a full ensemble analysis at higher resolution would be required to fully characterize the resolution dependence of our simulations. Furthermore, these simulations would need to use the higher resolution throughout the 205-kyr spin-up period in order to benefit from better-resolved bed topography. This is unfeasible with currently available computing resources.

Geophysical and terrestrial evidence consistent with re-advance. Previous geophysical and terrestrial observations are consistent with our proposed sequence of retreat and re-advance, but do not yet provide a coherent pattern of retreat and re-advance. Their spatial coverage is presently insufficient to reveal the full complexity of Holocene retreat and re-advance. For the Weddell Sea, ref. 53 presented evidence that Korff Ice Rise (KIR; Fig. 1) has been in a steady configuration since around 2.5 kyr BP. However, before that time KIR could have undergone substantial flow disturbance—including near-complete ungrounding and regrounding (as in our reference simulation; Supplementary Video 1)—if subsequent steady ice flow has had enough time to remove englacial evidence of such a flow disturbance. See ref. 53 for details of this interpretation. Radar data from BIR (Fig. 1) suggest a reorganization in flow as early as 4 kyr bp (ref. 10), while regional uplift rates suggest that BIR may have been ungrounded between 4 kyr BP and 2 kyr BP (ref. 5). In the Ellsworth 88 and Pensacola 89,90 Mountains, geological exposure-age dating techniques of the second sec niques constrain the thinning of the ice during the Holocene. These studies cannot provide evidence for re-thickening, which could be associated with re-advance, but the results cannot rule out the possibility of lowering of the ice-sheet surface below its present-day height and subsequent re-thickening within the last 4 kyr or $\mathrm{so}^{91}.$ Ref. 92 noted that radar-derived basal topography upstream of a subglacial basin beneath the Institute and Möller ice streams suggests a former grounding-line position more than 100 km upstream of today's grounding line, although these authors did not suggest that this was a Holocene grounding-line position.

Similarly, in the Ross Sea exposure-age dating in the Trans-Antarctic Mountains (see, for example, refs ^{3,4,40,93}) may be consistent with our conclusions, but cannot confirm or rule-out re-advance. Geophysical observations have hinted at recent re-advance. Borehole temperatures have been used to date the grounding and formation of Crary Ice Rise (CIR; Fig. 1b) to 1.5–1.0 kyr BP (ref. ⁵⁷) and ice-penetrating radar surveys of Kamb Ice Stream indicate that the grounding line was upstream of its present location during the past few centuries⁹. However, it is unclear whether the latter observation is evidence for a long-term large-scale re-advance, or for relatively-small-scale grounding-line fluctuations.

In both sectors, it is unclear whether these varied observations from diverse glacial environments (outlet glaciers, ice streams, ice rises and nunataks) paint a consistent picture of the timing of retreat and re-advance. Our work does not provide any detailed timing constraints; the timing of simulated grounding-line migration depends on uncertain bed topography and model parameters, and further work is needed to extract timing information from our radiocarbon and

radar observations. We leave to future work the important task of unravelling a retreat–readvance chronology consistent with all observations.

Code availability. The PISM code used in this study can be obtained from https://doi.org/10.5281/zenodo.1199066. Results and plotting scripts are available from https://doi.org/10.5880/PIK.2018.008. Scripts for processing and plotting radar data are also available on request.

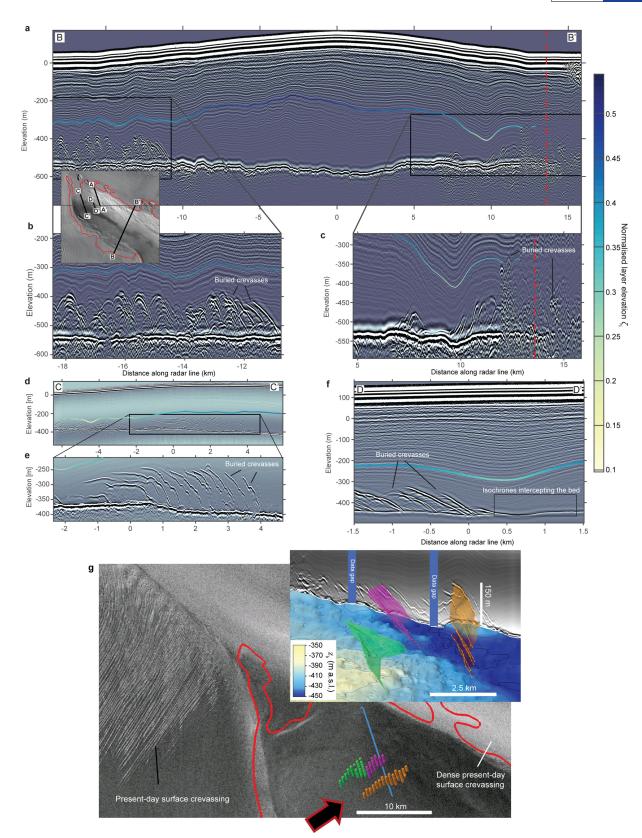
Data availability. Ice-penetrating radar data can be obtained from the UK Polar Data Centre at http://doi.org/99d. A simple MATLAB script for viewing the raw radar data is also provided at this link. The radiocarbon data supporting the findings of this study are available in Extended Data Table 1.

- Lipps, J. H., Ronan, T. DeLaca, T. Life below the Ross ice shelf. *Antarct. Sci.* 203, 447–449 (1979).
- Coenen, J. J. Inferring West Antarctic Subglacial Basin History and Ice Stream Processes Using Siliceous Microfossils. MSc Thesis, Northern Illinois Univ. (2016).
- Scherer, R. P. et al. Pleistocene collapse of the West Antarctic ice sheet. Science 281, 82–85 (1998).
- Abbott, M. B. & Stafford, T. W. Jr. Radiocarbon geochemistry of modern and ancient Arctic lake systems, Baffin Island, Canada. Quat. Res. 45, 300–311 (1996).
- 37. Priscu, J. C. et al. A microbiologically clean strategy for access to the Whillans lce Stream subglacial environment. *Antarct. Sci.* **25**, 637–647 (2013).
- Rosenheim, B. E. et al. Improving Antarctic sediment 14 C dating using ramped pyrolysis: an example from the Hugo Island trough. *Radiocarbon* 55, 115–126 (2013).
- 39. Člark, P. U. et al. The last glacial maximum. Science **325**, 710–714 (2009).
- Anderson, J. B. et al. Ross Sea paleo-ice sheet drainage and deglacial history during and since the LGM. Quat. Sci. Rev. 100, 31–54 (2014).
- Andrews, J. T. et al. Problems and possible solutions concerning radiocarbon dating of surface marine sediments, Ross Sea, Antarctica. *Quat. Res.* 52, 206–216 (1999).
- Licht, K. J. & Andrews, J. T. The ¹⁴C record of Late Pleistocene ice advance and retreat in the central Ross Sea, Antarctica. Arct. Antarct. Alp. Res. 34, 324–333 (2002).
- McKay, R. et al. Retreat history of the Ross Ice Sheet (Shelf) since the Last Glacial Maximum from deep-basin sediment cores around Ross Island. Palaeogeogr. Palaeoclimatol. Palaeoecol. 260, 245–261 (2008).
- Martinerie, P. et al. Physical and climatic parameters which influence the air content in polar ice. Earth Planet. Sci. Lett. 112, 1–13 (1992).
- Federer, U. et al. Continuous flow analysis of total organic carbon in polar ice cores. Environ. Sci. Technol. 42, 8039–8043 (2008).
- Antony, R. et al. Organic carbon in Antarctic snow: spatial trends and possible sources. *Environ. Sci. Technol.* 45, 9944–9950 (2011).
- Joughin, I. et al. Melting and freezing beneath the Ross ice streams, Antarctica. J. Glaciol. 50, 96–108 (2004).
- Christner, B. C. et al. A microbial ecosystem beneath the West Antarctic ice sheet. Nature 512, 310–313 (2014).
- Tulaczyk, S., Kamb, B. & Engelhardt, H. F. Estimates of effective stress beneath a modern West Antarctic ice stream from till preconsolidation and void ratio. *Boreas* 30, 101–114 (2001).
- Christianson, K. et al. Basal conditions at the grounding zone of Whillans Ice Stream, West Antarctica, from ice-penetrating radar. J. Geophys. Res. Earth Surf. 121, 1954–1983 (2016).
- Hall, B. L. et al. Constant Holocene Southern-Ocean 14 C reservoir ages and ice-shelf flow rates. Earth Planet. Sci. Lett. 296, 115–123 (2010).
- 52. Van Liefferinge, B. & Pattyn, F. Using ice-flow models to evaluate potential sites of million year-old ice in Antarctica. *Clim. Past* **9**, 2335 (2013).
- Kingslake, J. et al. Ice-flow reorganization in West Antarctica 2.5 kyr ago dated using radar-derived englacial flow velocities. Geophys. Res. Lett. 43, 9103–9112 (2016).
- Scambos, T., Haran, T., Fahnestock, M., Painter, T. & Bohlander, J. MODIS-based Mosaic of Antarctica (MOA) data sets: continent-wide surface morphology and snow grain size. *Remote Sens. Environ.* 111, 242–257 (2007).
- 55. Ely, J. et al. Insights on the formation of longitudinal surface structures on ice sheets from analysis of their spacing, spatial distribution, and relationship to ice thickness and flow. J. Geophys. Res. Earth Surf. 122, 961–972 (2017).
- Favier, L. & Pattyn, F. Antarctic ice rise formation, evolution, and stability. Geophys. Res. Lett. 42, 4456–4463 (2015).
- Bindschadler, R. A., Roberts, E. P. & Iken, A. Age of Crary Ice Rise, Antarctica, determined from temperature-depth profiles. Ann. Glaciol. 14, 13–16 (1990).
- Bueler, E. & Brown, J. Shallow shelf approximation as a "sliding law" in a thermomechanically coupled ice sheet model. J. Geophys. Res. Earth Surf. 114, F03008 (2009).
- The PISM authors. PISM, a Parallel Ice Sheet Model: user's manual (2017), based on development revision e9d2d1f8 (7 March 2017), http://www. pism-docs.org/wiki/lib/exe/fetch.php?media=pism_manual.pdf (2017).
- Cuffey, K. M. et al. Deglacial temperature history of West Antarctica. Proc. Natl Acad. Sci. USA 113, 14249–14254 (2016).
- Ligtenberg, S. et al. Future surface mass balance of the Antarctic ice sheet and its influence on sea level change, simulated by a regional atmospheric climate model. Clim. Dyn. 41, 867–884 (2013).
- Frieler, K. et al. Consistent evidence of increasing Antarctic accumulation with warming. Nat. Clim. Chang. 5, 348–352 (2015).
- Reese, R. et al. Antarctic sub-shelf melt rates via PICO. Cryosphere Discuss. https://doi.org/10.5194/tc-2017-70 (2017).



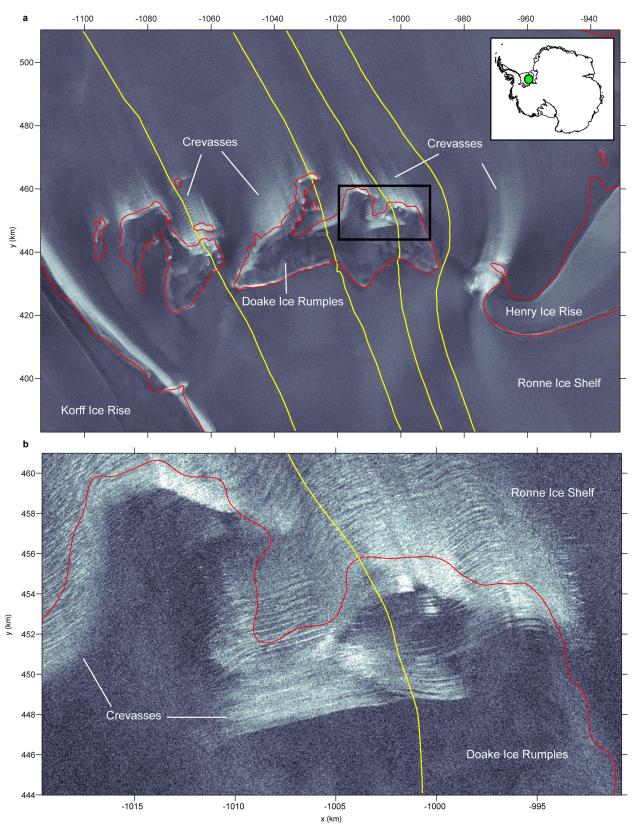
- Schmidtko, S. et al. Multidecadal warming of Antarctic waters. Science 346, 1227–1231 (2014).
- Li, C., von Storch, J. S. & Marotzke, J. Deep-ocean heat uptake and equilibrium climate response. Clim. Dyn. 40, 1071–1086 (2013).
- Levermann, A. et al. Kinematic first-order calving law implies potential for abrupt ice-shelf retreat. Cryosphere 6, 273–286 (2012).
- Pollard, D. & DeConto, R. Á. simple inverse method for the distribution of basal sliding coefficients under ice sheets, applied to Antarctica. *Cryosphere* 6, 953 (2012).
- Feldmann, J. et al. Resolution-dependent performance of grounding line motion in a shallow model compared with a full-Stokes model according to the MISMIP3d intercomparison. J. Glaciol. 60, 353–360 (2014).
- Stuhne, G. & Peltier, W. Reconciling the ICE-6G_C reconstruction of glacial chronology with ice sheet dynamics: the cases of Greenland and Antarctica. J. Geophys. Res. Earth Surf. 120, 1841–1865 (2015).
- Bueler, É., Lingle, C. S. & Brown, J. Fast computation of a viscoelastic deformable Earth model for ice-sheet simulations. *Ann. Glaciol.* 46, 97–105 (2007).
- Milne, G., Mitrovica, J. X. & Davis, J. L. Near-field hydro-isostasy: the implementation of a revised sea-level equation. *Geophys. J. Int.* 139, 464–482 (1999).
- Pritchard, H. D. Bedgap: where next for Antarctic subglacial mapping? Antarct. Sci. 26, 742–757 (2014).
- 73. Jones, P. W. First- and second-order conservative remapping schemes for grids in spherical coordinates. *Mon. Weath. Rev.* **127**, 2204–2210 (1999).
- Pollard, D. & DeConto, R. M. Description of a hybrid ice sheet-shelf model, and application to Antarctica. Geosci. Model Dev. 5, 1273–1295 (2012).
- Briggs, R. D., Pollard, D. & Tarasov, L. A data-constrained large ensemble analysis of Antarctic evolution since the Eemian. *Quat. Sci. Rev.* 103, 91–115 (2014).
- Pollard, D., Chang, W., Haran, M., Applegate, P. & DeConto, R. Large ensemble modeling of the last deglacial retreat of the West Antarctic Ice Sheet: comparison of simple and advanced statistical techniques. *Geosci. Mod. Dev.* 9, 1697–1723 (2016).
- Bindschadler, R. A. et al. Ice-sheet model sensitivities to environmental forcing and their use in projecting future sea level (the SeaRISE project). J. Glaciol. 59, 195–224 (2013).
- Lambeck, K. et al. Sea level and global ice volumes from the Last Glacial Maximum to the Holocene. Proc. Natl Acad. Sci. USA 111, 15296–15303 (2014).
- Bintanja, R. & Van de Wal, R. North American ice-sheet dynamics and the onset of 100,000-year glacial cycles. *Nature* 454, 869–872 (2008).
- Imbrie, J. D. & McIntyre, Á. SPECMAP time scale developed by Imbrie et al., 1984 based on normalized planktonic records (normalized O-18 vs time, specmap.017). Pangaea https://doi.org/10.1594/PANGAEA.441706 (2006).

- Gomez, N., Pollard, D. & Mitrovica, J. X. A 3-D coupled ice sheet-sea level model applied to Antarctica through the last 40 ky. *Earth Planet. Sci. Lett.* 384, 88–99 (2013).
- Whitehouse, P. L., Bentley, M. J., Milne, G. A., King, M. A. & Thomas, I. D. A new glacial isostatic model for Antarctica: calibrated and tested using observations of relative sea-level change and present-day uplift rates. *Geophys. J. Int.* 190, 1464–1482 (2012).
- 83. Jouzel, J. et al. Orbital and millennial Antarctic climate variability over the past 800,000 years. *Science* **317**, 793–796 (2007).
- Fudge, T. et al. Variable relationship between accumulation and temperature in West Antarctica for the past 31,000 years. *Geophys. Res. Lett.* 43, 3795–3803 (2016).
- 85. Hay, C. C. et al. Sea level fingerprints in a region of complex Earth structure: the case of WAIS. *J. Clim.* **30**, 1881–1892 (2017).
- Ji, F. et al. Variations of the effective elastic thickness over the Ross Sea and Transantarctic Mountains and implications for their structure and tectonics. *Tectonophysics* 717, 127–138 (2017).
- 87. Chen, B., Haeger, C., Kaban, M. K. & Petrunin, A. G. Variations of the effective elastic thickness reveal tectonic fragmentation of the Antarctic lithosphere. *Tectonophysics* https://doi.org/10.1016/j.tecto.2017.06.012 (2017).
- Hein, Á. S. et al. Mid-Holocene pulse of thinning in the Weddell Sea sector of the West Antarctic ice sheet. Nat. Commun. 7, 12511 (2016).
- Balco, G. et al. Cosmogenic-nuclide exposure ages from the Pensacola Mountains adjacent to the Foundation Ice Stream, Antarctica. Am. J. Sci. 316, 542–577 (2016).
- Bentley, M. J. et al. Deglacial history of the Pensacola Mountains, Antarctica from glacial geomorphology and cosmogenic nuclide surface exposure dating. *Quat. Sci. Rev.* 158, 58–76 (2017).
- Whitehouse, P. L. et al. Controls on Last Glacial Maximum ice extent in the Weddell Sea embayment, Antarctica. J. Geophys. Res. Earth Surf. 122, 371–397 (2017).
- Ross, N. et al. Steep reverse bed slope at the grounding line of the Weddell Sea sector in West Antarctica. Nat. Geosci. 5, 393 (2012).
- Todd, C., Stone, J., Conway, H., Hall, B. & Bromley, G. Late Quaternary evolution of Reedy Glacier, Antarctica. Quat. Sci. Rev. 29, 1328–1341 (2010).
- Jezek, K. C., Curlander, J. C., Carsey, F., Wales, C & Barry, R. RAMP AMM-1 SAR image mosaic of Antarctica, version 2. National Snow and Ice Data Center https://doi.org/10.5067/8AF4ZRPULS4H (2013).
- Fürst, J. J. et al. The safety band of Antarctic ice shelves. Nat. Clim. Chang. 6, 479–482 (2016).
- WAIS Divide Project Members. Precise interpolar phasing of abrupt climate change during the last ice age. Nature 520, 661–665 (2015).



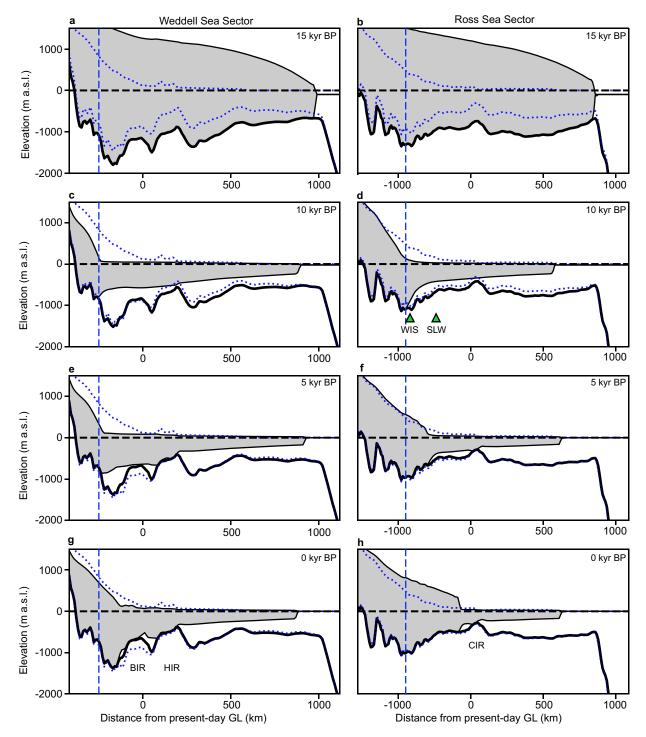
Extended Data Fig. 1 | **Relic crevasses in Henry Ice Rise. a**, Radargram, aligned perpendicular to the divide ridge (inset shows the location). One undulating isochrone is delineated with colours showing normalized elevation. **b**, **c**, Close-up views of the boxed regions indicated in **a**. In both close-up panels, diffractors (hyperbolic reflectors) are interpreted as expressions of relic crevasses (data are unmigrated). The red vertical dashed line is the present-day grounding line³¹. **d**-**f**, Radargrams aligned approximately perpendicular to northern relic crevasses (**d** and **e** show

migrated data). In c (6 km \leq x \leq 8 km) and f (0.3 km \leq x \leq 1.4 km) isochrones intercepting the bed are evident. g, Three relic crevasses mapped across several radar lines over a Radarsat Antarctic Mapping Project (RAMP) image⁹⁴. The inset is an oblique, three-dimensional view of the features over an interpolated surface, showing the bed elevation z_b (see Methods). Crevasse spacing in these areas ranges between approximately 200 m and 600 m. The arrow indicates the view direction of the oblique view.



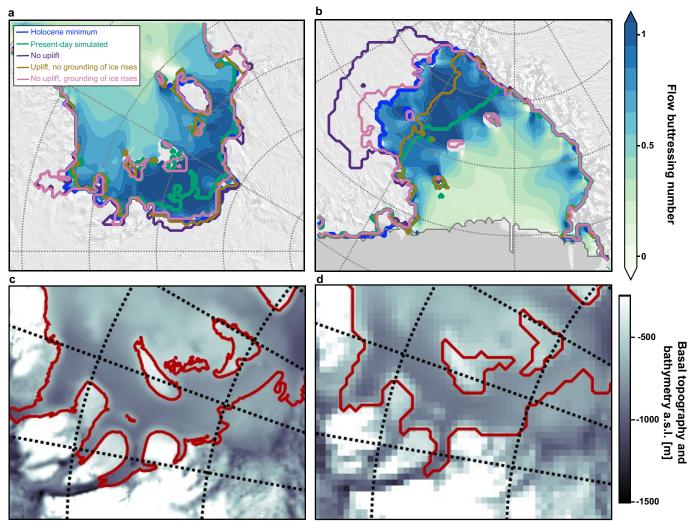
Extended Data Fig. 2 | Crevassing at Doake Ice Rumples. a, RAMP⁹⁴ image showing the surface expression of ice-shelf crevasses in synthetic aperture radar data. Light areas indicate high backscatter from (near-) surface reflectors, interpreted to be surface crevasses. Crevasses form over and immediately downstream of Doake Ice Rumples. We hypothesize that crevasses once formed in a similar manner over the topographic high beneath the northern tip of HIR. b, Close-up view of the crevasses (the black box in a shows the location), whose spacing (100–300 m),

orientation (perpendicular to the flow of the ice shelf) and lateral extent (roughly 10 km) are similar to the steeply dipping reflectors discovered near the bed of the northern tip of HIR (for example, Extended Data Fig. 2g) in the region of a topographic high. Yellow curves are flow lines computed from satellite-derived surface velocities³⁰. Flow is from bottom to top. Polar stereographic coordinates are in km. The present-day grounding line³¹ is in red.



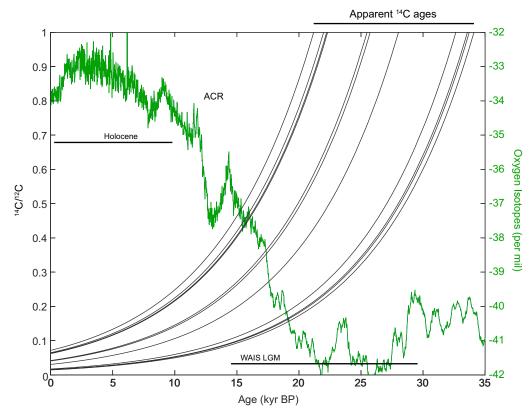
Extended Data Fig. 3 | Modelled grounding-line retreat and lithospheric rebound. Cross-sections along transects through the Weddell (left) and Ross (right) Sea sectors, at 5-kyr intervals (for transect locations, see Fig. 3). The horizontal axis shows the distance from the present-day grounding line. The vertical blue dashed line shows the position of maximum grounding-line retreat. a, b, 15 kyr BP, with the grounding line close to the continental shelf edge. c, d, 10 kyr BP, with the grounding line having retreated to approximately its minimum, most retreated location. e, f, 5 kyr BP, with both ice shelves grounded on sub-ice-shelf bathymetric

highs owing to seafloor uplift. **g**, **h**, Present day, with the grounding line having re-advanced to roughly the present-day configuration in response to the grounding of the ice shelf and uplift at the grounding line. The Crary, Bungenstock and Henry ice rises (CIR, BIR and HIR) are labelled in **g** and **h**. The Whillans Ice Stream (WIS) and Subglacial Lake Whillans (SLW) sediment-core locations are labelled in **d**. Blue dotted lines show the observed present-day ice-sheet bed, ocean floor and ice surface²⁹, remapped on to the 15-km grid of the ice-sheet model.



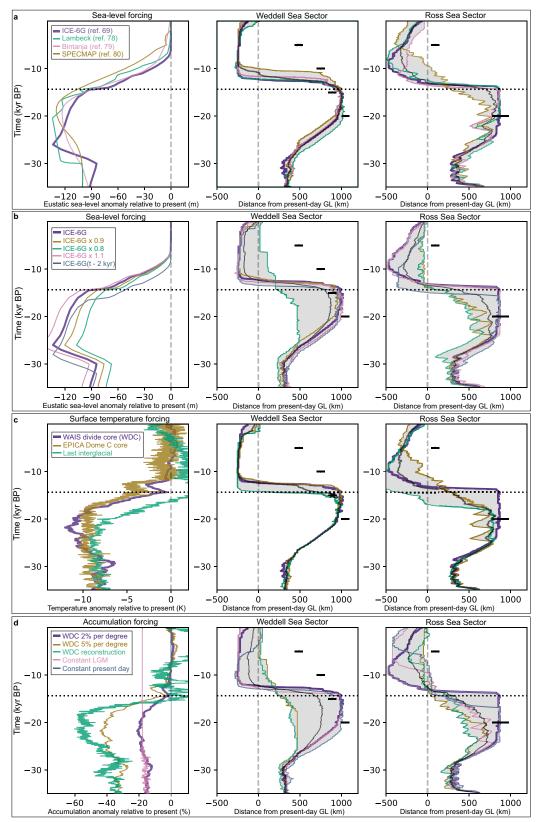
Extended Data Fig. 4 | Drivers of re-advance and the impact of bed re-mapping. a, b, Results from four simulations (the reference simulation, and three additional experiments) designed to examine the cause of re-advance in the Weddell Sea (a) and Ross Sea (b) sectors (Methods). The most inland grounding-line location in the reference simulation, at around 10 kyr BP, is in blue. The colour map shows the flow buttressing number 95 at 10 kyr BP in the 'No uplift, grounding of ice rises' experiment. The ice-front position is in grey. Background images over the grounded ice sheet

are from MOA³². **c**, Basal topography and bathymetry in the Weddell Sea sector (with the grounding line in red) according to a 1-km-resolution dataset, constrained by geophysical observations (Bedmap 2; ref. ²⁹). **d**, Conservative remapping of these data to 15-km resolution. Remapping substantially lowers the apparent maximum bed elevations beneath ice rises in the Weddell Sea sector: 135 m at KIR, 112 m at HIR and 36 m at BIR.



Extended Data Fig. 5 | True and apparent ages of radiocarbon. The 11 grey lines show exponential $^{14}\mathrm{C}$ -decay curves connecting the $^{14}\mathrm{C}/^{12}\mathrm{C}$ ratios (scale on the left) measured on acid-insoluble inorganics (AIOs) from our subglacial sediment samples to the apparent radiocarbon ages calculated from these measurements. The latter calculation assumes that the initial $^{14}\mathrm{C}/^{12}\mathrm{C}$ ratios in AIOs were equal to the modern ratio in radiocarbon dating standards. As discussed in the text and Methods,

organic matter in Antarctic glacigenic sediments frequently contains an admixture of old ¹⁴C-dead material^{41,42}. The record of oxygen isotopes in water ice from the WAIS Divide ice core (green line, with scale on the right) provides climatic context for the period between now and 35 kyr BP (ref. ⁹⁶). Three key climatic periods are labelled: WAIS LGM³⁹, Antarctic cold reversal (ACR) and Holocene.

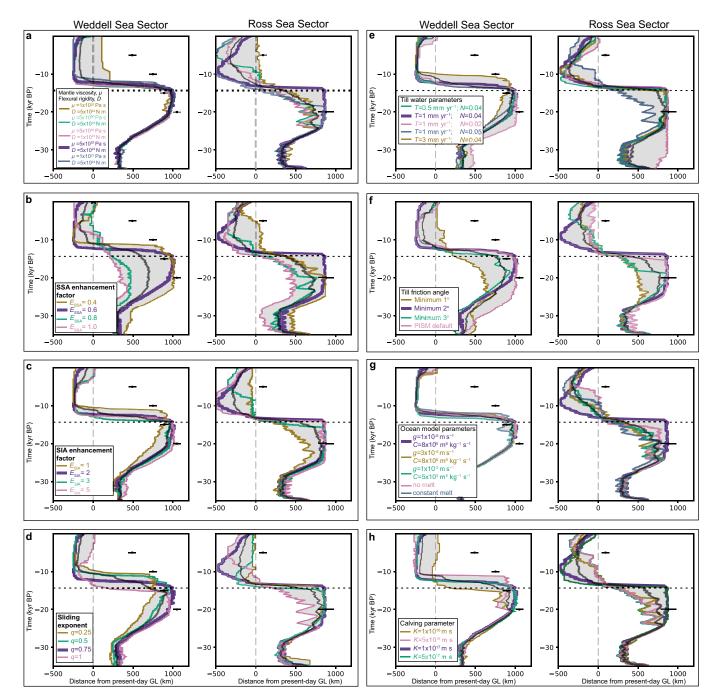


Extended Data Fig. 6 | See next page for caption.



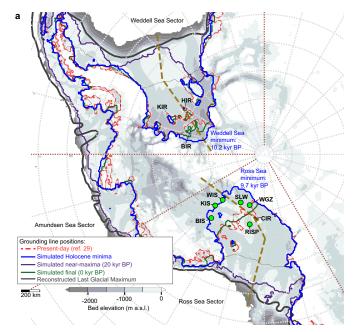
Extended Data Fig. 6 | Model sensitivity to forcings. In the middle and right panels are time series of grounding-line position along transects, showing model sensitivity in the Weddell Sea (middle panels) and Ross Sea (right panels) sectors to: a, different sea-level reconstructions^{69,78-80}; b, different scalings of the sea-level forcing to mimic self-gravitational effects; c, different surface-temperature forcings; and d, different accumulation forcings. In the left panels are: a, four alternative sea-level reconstructions; b, three alternative scalings of the reference-simulation sea-level forcing and a version that has been uniformly shifted 2,000 years earlier; c, temperature reconstructions from two ice cores, WAIS Divide and EPICA Dome C (EDC), and a reconstruction from the Last Interglacial (from EDC data); and d, four alternative accumulation histories. The constant LGM accumulation uses the EPICA Dome C

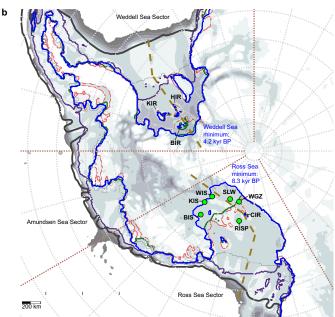
core⁸³ and a scaling of 2% per degree. Temperature and accumulation are expressed relative to the present day. Grounding-line positions are relative to the present-day position (vertical dashed line) along the transacts shown in Fig. 3. In all simulations, the grounding line is in its most advanced position, up to 1,000 km beyond its present-day position, before MWP1a (14.4 kyr BP; horizontal dotted line). During the Holocene the grounding line retreats up to 500 km upstream of its present location, and usually re-advances towards its present-day position. Grey shading indicates the spread of grounding-line responses, and grey curves show the mean of each sensitivity experiment. In each case the violet curve shows the reference simulation. Grounding-line positions (based on marine and terrestrial geological evidence) from the RAISED reconstruction with associated uncertainties are shown in black².



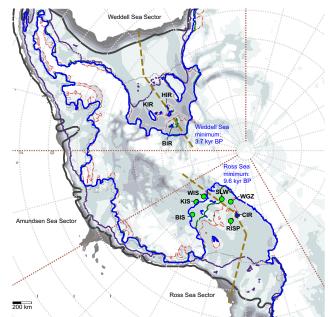
Extended Data Fig. 7 | **Model sensitivity to parameters.** Time series of grounding-line position along transects, showing model sensitivity in the Weddell and Ross Sea sectors to: **a**, mantle viscosity, μ , and the flexural rigidity of the lithosphere, D; **b**, **c**, enhancement factors $E_{\rm SSA}$ and $E_{\rm SIA}$; **d**, the sliding-law exponent, q; **e**, the till water decay rate, T, and till effective pressure fraction, N; **f**, minimum till friction angle and the method used to derive the friction angle (see Methods); **g** PICO ocean

model parameters for overturning strength, C, and heat exchange, g; and \mathbf{h} , the dependence of calving rate on the ice-shelf spreading rate, K (Extended Data Table 2). In each panel, the violet curve shows the reference simulation. Grounding-line positions (based on marine and terrestrial geological evidence) from the RAISED reconstruction, with associated uncertainties, are shown in black².





Extended Data Fig. 8 | Model sensitivity to spatial resolution. The results of three simulations using different grid resolutions: a, 15 km (reference simulation; identical to Fig. 3); b, 10 km; and c, 7 km. Owing to computational limitations, the two higher-resolution simulations cover only the past 20 kyr, so they lack a higher-resolution spin-up period. These higher-resolution simulations display similar Holocene retreat and



re-advance driven by isostatic rebound to the reference simulation, but the LGM extent and grounding-line re-advance in the Weddell Sea are much less. A full exploration of the resolution dependence of the model requires using higher resolution during entire simulations for all ensemble members. This is limited at present by computing resources. Background shading shows basal topography and bathymetry²⁹.



Extended Data Table 1 \mid Results of radiocarbon and $\delta^{13}C$ analyses of subglacial sediments

K-Keck# Ua-					
Uppsala# K-166508	Sample Name RISP Core 5 0-2 cm	Fraction Modern 0.0807	¹⁴ C Age (BP) 20,220	± 60	δ ¹³ C ‰ -28.2
K-166509	RISP Core 5 12-15 cm	0.0351	26,920	150	-20.2
K-166510	RISP Core 5 28.5-31.5 cm	0.0397	25,920	100	-27.0
K-166511	RISP Core 5 51-54 cm	0.0409	25,690	90	-26.2
K-166512	RISP Core 12 0-2 cm	0.1058	18,050	100	-20.2
K-166513	RISP Core 12 117-118cm	0.0604	22,550	70	-27.6
K-100313 K-154083	WGZ-1 SS22 net bulk	0.0612	22,440	180	-27.3
K-160518	WGZ-1-MC1A-1-2cm	0.0809	20,205	50	-24.3
K-160516 K-160516	WGZ-1-MC5A-3-5cm	0.0676	21,650	60	-24.5 -24.1
K-160510 K-160517	WGZ-1-MC5A-15-17cm	0.0424	25,400	80	-24.1
K-166502	WGZ-1-MC3A-13-17cm	0.0424	29,950	220	-24.0 -27.7
K-166503	WGZ-1 GC-4 36-38 cm	0.0240	30,930	240	-26.6
K-166504	WGZ-1 GC-4 45-47 cm	0.0541	23,440	120	-25.6
K-160504 K-160514	WGZ-1-GC-4-13-16cm	0.0416	25,540	90	-25.9
K-160514 K-160515	WGZ-1-GC-4-13-16cm	0.0562	23,120	90 70	-25.9 -25.6
K-160313 K-154080	WGZ-1-GC-4-51-556111 WGZ-1 GP1-top SS-8	0.0582	22,870	260	-23.8
K-154080 K-154081	WGZ-1 GP1-top SS-8	0.0576	22,930	190	-23.0
K-154081 K-154082	WGZ-1 GC-1 SS-7 86	0.0370	24,530	380	-25.3
K-154062 K-160509	WGZ-1-GC-1-9-12cm	0.0472	25,220	100	-25.3 -25.2
K-160509 K-160510	WGZ-1-GC-1-38-43cm	0.0498	24,100	80	-23.9
K-160510 K-160511	WGZ-1-GC-1-58-43CIII WGZ-1-GC-1-52-56cm	0.0498	25,140	140	-23.9 -24.6
K-160511 K-160512	WGZ-1-GC-1-65-68cm	0.0437	21,400	60	-24.0
K-160512 K-160513	WGZ-1-GC-1-80-84cm	0.0441	25,070	90	-24.5
Ua-51811	WGZ-1-GC-1-80-84cm	0.0441	28,687	161	-23.0 -24 ⁹
K-160521	SLW-PEC-1-34-35cm	0.0648	21,990	70	-24.6
K-154084	SLW-1 MC1B 0-8 bulk	0.0306	28,020	230	-24.9
Ua-51810	SLW-1, MC1B, 30-31 cm	-	29,378	180	-24.9
K-166505	UpB 89-4-50-53 cm	0.0406	25,740	100	-25,1
K-166506	UpB 89-7-50 cm	0.0419	25,490	90	-25.6
K-160500	UpB-88-89-SampleB2	0.0627	22,240	60	-26.1
K-160519	UpB-91-92-12-22-91-TV	0.0715	21,190	70	-25.4
K-166507	UpB 95-3-1-1	0.0621	22,320	70 70	-20.4
K-166499	KIS 96-12-1-2-2-2 top 10cm	0.0150	33,720	230	-26.8
K-166501	KIS 96-7-1-3	0.0143	34,100	240	-26.7
K-166500	KIS 00-5-1-1C	0.0143	33,570	230	-26.7 -26.7
K-166498	BIS 98-2-2-3c 60-70cm	0.0172	32,640	210	-26.5
K-154085	WGZ modern amphipod1	0.8738	1,085	20	-20.5
K-154086	WGZ modern amphipod2	0.8746	1,005	15	= =
K-154087	WGZ modern amphipod3	0.8669	1,145	15	_
10-10-1007	VV GZ modem ampriipodo	0.0000	1,170	10	-

Carbon-isotope results, including the fraction of modern carbon, calculated age, analytical error and independently measured δ^{13} C value. A low fraction of modern carbon relative to dominant ancient (radiocarbon-dead) carbon skews apparent ages older than the actual age of the marine connection discussed here. The light δ^{13} C results also point to a substantial source of old carbon. UpB is the upstream portion of the WIS.



Extended Data Table 2 | Key model parameters, with modelled retreat and re-advance

Parameter [Unit], symbol	Reference	Range	Relevance for LGM extent, overshoot of the present-day GL position during retreat and re-advance of the GL
Mantle viscosity [Pa s], μ	5x10 ²⁰	10 ²⁰ -10 ²¹	more overshoot retreat for higher values but delayed re-advance
Flexural rigidity [N m], D	5x10 ²⁴	5×10 ²³ -5×10 ²⁴	smaller values delay retreat and re-advance
SSA enhancement [], E _{SSA}	0.6	0.4-0.8	higher values cause less extended LGM state and less overshoot
SIA enhancement [], $E_{\rm SIA}$	2	1-5	smaller values cause less extended LGM state and less overshoot
Flow law exponent [], q	0.75	0.25-1	smaller values cause less extended LGM state and delayed retreat
Till water decay rate [mm yr $^{-1}$], T	3	0.5-3	higher values cause slightly less extended LGM
Till effect. overburden [], N	0.04	0.02-0.05	higher values cause slightly less extended LGM
Till friction angle min [°], ϕ	2	1-3	relevant for LGM state and hence for overshoot retreat and re-advance
PICO heat exchange [m s ⁻¹], g	1x10 ⁻⁵	1-3x10 ⁻⁵	2 nd order influence on LGM state and timing of retreat
PICO overturning [m ⁶ kg ⁻¹ s ⁻¹], C	8x10 ⁵	8-20x10 ⁵	2 nd order influence on LGM state and timing of retreat
Eigen calving [m s], K	1x10 ¹⁷	1x10 ¹⁶ -5x10 ¹⁷	relevant for LGM state and initiation of retreat via ice-shelf buttressing

The table shows key model parameters that have been varied as part of our sensitivity study of our ice-sheet model (see Methods). In each case, the value used in the reference simulation is given, as well as the range over which the parameters were varied during the sensitivity study. Also provided is a summary of the impact of each parameter on the model behaviour with respect to the retreat of the grounding line past its present-day location and its subsequent re-advance. See Methods for a detailed discussion of model sensitivities.



Quantitative phosphoproteomic analysis of the molecular substrates of sleep need

Zhiqiang Wang¹, Jing Ma¹, Chika Miyoshi¹, Yuxin Li², Makito Sato¹, Yukino Ogawa¹, Tingting Lou¹, Chengyuan Ma³, Xue Gao³, Chiyu Lee¹, Tomoyuki Fujiyama¹, Xiaojie Yang¹, Shuang Zhou³, Noriko Hotta-Hirashima¹, Daniela Klewe-Nebenius¹, Aya Ikkyu¹, Miyo Kakizaki¹, Satomi Kanno¹, Liqin Cao¹, Satoru Takahashi⁴, Junmin Peng², Yonghao Yu⁵, Hiromasa Funato^{1,6*}, Masashi Yanagisawa^{1,7,8*} & Qinghua Liu^{1,3,9,10*}

Sleep and wake have global effects on brain physiology, from molecular changes¹⁻⁴ and neuronal activities to synaptic plasticity³⁻⁷. Sleep-wake homeostasis is maintained by the generation of a sleep need that accumulates during waking and dissipates during sleep⁸⁻¹¹. Here we investigate the molecular basis of sleep need using quantitative phosphoproteomic analysis of the sleep-deprived and Sleepy mouse models of increased sleep need. Sleep deprivation induces cumulative phosphorylation of the brain proteome, which dissipates during sleep. Sleepy mice, owing to a gain-of-function mutation in the Sik3 gene¹², have a constitutively high sleep need despite increased sleep amount. The brain proteome of these mice exhibits hyperphosphorylation, similar to that seen in the brain of sleep-deprived mice. Comparison of the two models identifies 80 mostly synaptic sleep-need-index phosphoproteins (SNIPPs), in which phosphorylation states closely parallel changes of sleep need. SLEEPY, the mutant SIK3 protein, preferentially associates with and phosphorylates SNIPPs. Inhibition of SIK3 activity reduces phosphorylation of SNIPPs and slow wave activity during non-rapid-eye-movement sleep, the best known measurable index of sleep need, in both Sleepy mice and sleep-deprived wildtype mice. Our results suggest that phosphorylation of SNIPPs accumulates and dissipates in relation to sleep need, and therefore SNIPP phosphorylation is a molecular signature of sleep need. Whereas waking encodes memories by potentiating synapses, sleep consolidates memories and restores synaptic homeostasis by globally downscaling excitatory synapses⁴⁻⁶. Thus, the phosphorylationdephosphorylation cycle of SNIPPs may represent a major regulatory mechanism that underlies both synaptic homeostasis and sleep-wake homeostasis.

Homeostatic sleep regulation is a global, intrinsic and cumulative process that ultimately involves most brain cells and regions^{3,5,7}; this is distinct from executive switching between sleep and wake states, which is controlled by specific neural circuits^{13,14}. We hypothesize that the molecular substrates of sleep need satisfy four criteria: 1) they should be globally and similarly regulated in most brain cells or regions; 2) they should accumulate gradually during waking and dissipate through sleep; 3) they should change in parallel with sleep need in different contexts; and 4) gain or loss of these functions should cause bidirectional changes of sleep need.

Sleep deprivation increases sleep need in mice, as shown by enhanced slow wave activity (SWA) or delta power (1–4 Hz) of electroencephalography (EEG) during non-rapid-eye-movement sleep (NREMS), which declines rapidly to the baseline followed by rebound sleep in early dark phase^{8,10,12} (Fig. 1a, Extended Data Fig. 1a–e). We recently identified a dominant mutation in *Sleepy* mice¹², *Sik3*^{Slp/+}, in which a single

nucleotide substitution in the gene for salt-inducible kinase 3 (SIK3), a member of the AMP-activated protein kinase (AMPK) family¹⁵, causes constitutively high sleep need, manifested by elevated SWA and duration of NREMS (Extended Data Fig. 1f–i). Sleep deprivation increases wake time, whereas the *Sleepy* mutation decreases wake time; yet both cause elevated sleep need. We hypothesized that cross-comparison of these contrasting models of increased sleep need would reveal specific molecular changes associated with sleep need by filtering out non-specific effects of prolonged sleep, wake and stress.

We subjected three groups of wild-type C57BL/6N mice, at Zeitgeber time (ZT) zero, to 6 h of ad libitum sleep (S6) or sleep deprivation (SD6), or 6 h of sleep deprivation followed by a 3-h recovery sleep (RS3), respectively (Fig. 1a). We collected brains from wild-type $(Sik3^{+/+})$ and Sleepy $(Sik3^{Slp/+})$ mice at ZT12.5, the lowest point of SWA in wild-type mice (Fig. 1a). As shown by immunoblotting with antibodies against 14 phosphorylated substrate motifs, global phosphorylation of substrates of AMPK, protein kinase C (PKC), protein kinase A (PKA) and 'ataxia telangiectasia mutated' (ATM) and 'ATM and RAD3-related' (ATR) kinases was specifically increased in brains of both Sleepy mice and wild-type SD6 mice, but was not affected by fasting (Fig. 1b, c and Extended Data Figs. 2, 3). By contrast, other signalling pathways, such as casein kinase II (CK2) or tyrosine kinases, were not significantly affected (Fig. 1c and Extended Data Fig. 2). These observations indicate that similar kinase pathways are globally activated in Sleepy and sleep-deprived brains.

Next, we performed quantitative proteomic and phosphoproteomic studies of whole brain lysates using multiplex tandem mass tag (TMT) labelling coupled with liquid chromatography-mass spectrometry (LC-MS)¹⁶⁻¹⁹ (Fig. 1a). A total of 4 proteomic and 13 phosphoproteomic experiments was performed (Supplementary Tables 1, 2). The amount of peptides or phosphopeptides corresponding to exon 13, which is not translated in the Sik3^{Slp} mutant allele, was specifically reduced by 40% in Sik3^{Slp/+} relative to Sik3^{+/+} samples (Fig. 1d, g and Extended Data Fig. 4a); this acted as a stringent internal control. In summary, brain proteomic analysis quantified 7,963 proteins, of which 5,280 were present in all conditions in the pairwise comparisons of SD6 and RS3 (5,769 proteins), SD6 and S6 (6,067 proteins), and Sleepy and wild-type (7,650 proteins) groups (Extended Data Fig. 4b-h, Supplementary Table 1). Phosphoproteomic analysis quantified a total of 62,384 unique phosphopeptides from 7,104 phosphoproteins and identified 51,821 phosphorylation sites (Supplementary Table 2a).

Few quantified peptides or proteins showed significant changes in abundance (Q < 0.2) in the comparisons of brain proteomes between *Sleepy* and wild-type (0.09%; 3.5%), SD6 and RS3 (0.01%; 0%), or SD6 and S6 (0%; 0.01%) samples (Fig. 1d–f and Extended Data

¹International Institute for Integrative Sleep Medicine (WPI-IIIS), University of Tsukuba, Tsukuba, Japan. ²Departments of Structural Biology and Developmental Neurobiology, St. Jude Proteomics Facility, St. Jude Children's Research Hospital, Memphis, TN, USA. ³National Institute of Biological Sciences, Beijing, China. ⁴Laboratory Animal Resource Center, University of Tsukuba, Tsukuba, Japan. ⁵Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX, USA. ⁶Department of Anatomy, Faculty of Medicine, Toho University, Tokyo, Japan. ⁷Department of Molecular Genetics, University of Texas Southwestern Medical Center, Dallas, TX, USA. ⁸Life Science Center for Survival Dynamics (TARA), University of Tsukuba, Tsukuba, Japan. ⁹Tsinghua Institute of Multidisciplinary Biomedical Research, Tsinghua University, Beijing, China. ¹⁰Department of Biochemistry, Department of Neuroscience, Center for Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX, USA. *e-mail: funato.hiromasa.km@u.tsukuba.ac.jp; yanagisawa.masa.fu@u.tsukuba.ac.jp; Qinghua.liu@utsouthwestern.edu

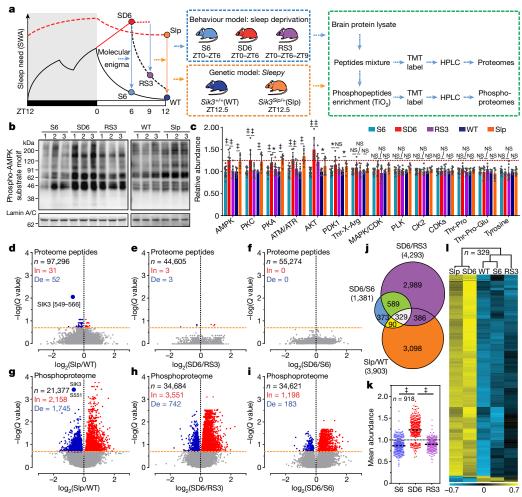


Fig. 1 | Brains of *Sleepy* mutant mice exhibit hyperphosphorylation that mimics that in sleep-deprived brains. a, Experimental design for proteomic and phosphoproteomic analysis of two models. b, Representative immunoblots of brain lysates with antibody specific for AMPK target phosphorylation motifs. Blots represent three (sleep-deprived) or two (*Sleepy*) independent experiments. c, Quantitative analysis of immunoblots using specific antibodies for different phosphorylated protein motifs. n = 12 (S6), 9 (SD6, RS3), 6 (wild-type (WT), *Sleepy* (Slp)). Mean \pm s.d., two-way ANOVA, Fisher's least significant difference (LSD). d–i, Volcano plots showing changes in

peptides (**d**-**f**) and phosphopeptides (**g**-**i**) in *Sleepy*/wild-type, SD6/RS3 and SD6/S6 comparisons. Multiple unpaired t-test (P value) followed by false discovery rate (FDR) (Q value) analysis. In, increase; De, decrease. **j**, Venn diagram of significantly changed phosphopeptides among three groups, with the number of significantly changed phosphopeptides in each experiment shown in parentheses. **k**, Analysis of mean abundance of 918 phosphopeptides that are changed in both SD6/RS3 and SD6/S6 comparisons. Mean, one-way ANOVA, Dunnett's test. **l**, Hierarchical cluster analysis of 329 phosphopeptides that are changed in all three groups. *P<0.05; ‡P<0.001; NS, not significant (P>0.05).

Fig. 4g), suggesting that the whole brain proteome was globally stable (Supplementary Discussion 1). By contrast, comparison of the brain phosphoproteomes showed that a sizable portion of phosphopeptides exhibited significant changes (Q < 0.2) between the SD6 and RS3 (12.4%), SD6 and S6 (4%), and *Sleepy* and wild-type (18.3%) conditions (Fig. 1g-j). In sleep-deprived brains, the majority of changes in phosphorylation are increases: SD6/RS3 (3,551/4,293, 82.7%) or SD6/S6 (1,198/1,381, 86.7%) (Fig. 1h, i). The mean abundance of 918 phosphopeptides that were changed in both SD6/RS3 and SD6/S6 groups was around 32% or around 25% lower in S6 or RS3 brains, respectively, than in SD6 brains (Fig. 1j, k). This asymmetric increase in phosphorylation was not observed in the liver phosphoproteome after sleep deprivation (Extended Data Fig. 5). Instead, the liver phosphoproteome showed decreases in global phosphorylation in these comparisons: SD6/S6 (1,275/2,186, 58.3%) and SD6/RS3 (286/433, 66.1%) (Extended Data Fig. 5b, c). These studies suggest that sleep and wake have opposing effects on the brain phosphoproteome: prolonged wakefulness causes hyperphosphorylation, whereas sleep promotes global dephosphorylation of the brain proteome.

Comparison of *Sleepy* and sleep-deprived models reveals 329 phosphopeptides that are significantly (Q < 0.2) altered in all three (*Sleepy*/

wild-type, SD6/S6 and SD6/RS3) comparisons (Fig. 1j). On the basis of the mean abundance of each of these 329 phosphopeptides, unsupervised cluster analysis groups *Sleepy* samples with SD6 samples, whereas wild-type samples cluster with S6 and RS3 samples (Fig. 1l). We used antibodies against specific phosphorylation sites to confirm hyperphosphorylation of multiple proteins in both *Sleepy* and SD6 samples (Extended Data Fig. 4i, j). These results suggest that *Sleepy* mutant brains exhibit a global hyperphosphorylation of proteins, mimicking that seen in sleep-deprived brains.

Protein functions can be switched on or off by site-specific phosphorylation, or modulated by cumulative phosphorylation of multiple sites $^{20-23}$. We noted a group of proteins containing multiple phosphorylation sites that appear to be co-ordinately regulated in both *Sleepy* and SD6 models (Extended Data Fig. 6a, b). For example, the synaptic vesicle protein synapsin-1 contains multiple functionally important phosphorylation sites 21,22 , almost all of which are hyperphosphorylated in brains of sleep-deprived or *Sleepy* mice (Fig. 2a and Extended Data Fig. 6a). We measured overall phosphorylation state change (Δ Ps) of synapsin-1 by calculating the sum of $\log_2(\text{fold change})$ values of all significantly (Q < 0.2) changed synapsin-1 phosphopeptides. Synapsin-1 has Δ Ps values of 7.5, 5.5 and 13.7 in the SD6/RS3,

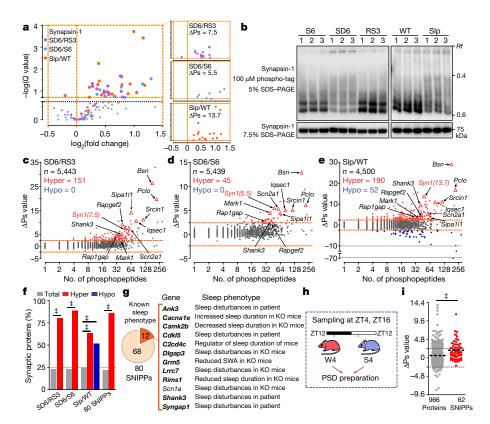


Fig. 2 | Changes in phosphorylation state of SNIPPs parallel changes in sleep need.

a, Volcano plots of quantified phosphopeptides of synapsin-1 in SD6/RS3 (violet), SD6/S6 (blue) and Sleepy/wild-type (orange) comparisons. Multiple unpaired t-test (P value) followed by FDR (Q value) analysis. b, Phosphorylation of synapsin-1 was assessed by regular or phospho-tag SDS-PAGE followed by immunoblotting (two independent experiments). c-e, Global ∆Ps analysis of phosphoproteins in three comparisons. Numbers of hyperphosphorylated (Hyper) and hypophosphorylated (Hypo) peptides in each comparison are shown. Labels show genes encoding the proteins. Dotted lines, $\Delta Ps = \pm 2.4$. f, Percentage of synaptic proteins in total, hypophosphorylated and hyperphosphorylated proteins, and among 80 SNIPPs. χ^2 test, two-sided. **g**, Mutations in 12 SNIPP genes cause sleep phenotypes. Genes for synaptic proteins are shown in bold. h, A schematic of the normal sleep-wake model⁴. i, Quantitative ΔPs analysis of SNIPPs in the W4/S4 model. n = 966 (total), 62 (SNIPPs). Mean; unpaired *t*-test, two-tailed. $\ddagger P < 0.001$.

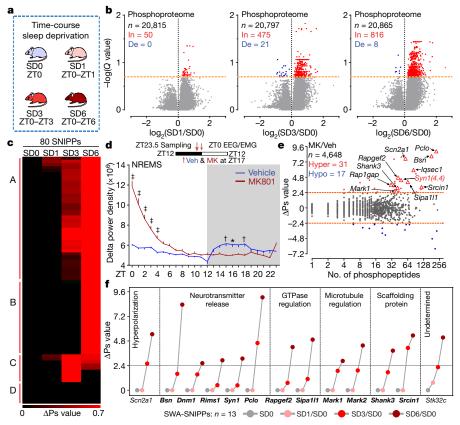


Fig. 3 | **SNIPPs exhibit time-dependent cumulative phosphorylation. a**, A schematic of time-course sleep deprivation. **b**, Volcano plots comparing phosphoproteomes of SD1/SD0, SD3/SD0 and SD6/SD0. Multiple unpaired *t*-test (*P* value) followed by FDR (*Q* value) analysis. **c**, Temporal profile and classification of phosphorylation-state changes of SNIPPs. **d**, Circadian analysis of absolute NREMS delta power of

vehicle (Veh) or MK801 (MK)-injected mice (n=14). Mean \pm s.e.m., two-way ANOVA, Sidak's test *P < 0.05; †P < 0.01; ‡P < 0.001. e, Global Δ Ps analysis of MK801/vehicle group. f, Time-dependent cumulative phosphorylation of 13 SWA-SNIPPs that occur in *Sleepy*, SD and MK801 models. Genes encoding synaptic proteins are shown in bold.

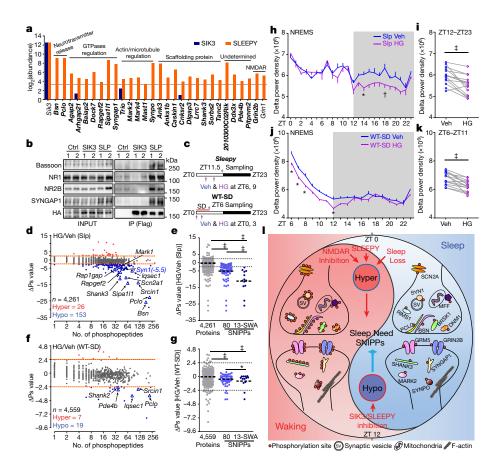


Fig. 4 | SLEEPY preferentially interacts with SNIPPs and alters sleep-wake homeostasis. a, Comparison of mass-spectrometry signals of SNIPPs in immunoprecipitates of SLEEPY and SIK3. b, Immunoprecipitation confirms interactions between SLEEPY and SNIPPs (two independent experiments). c, A schematic of SIK3 inhibition in Sleepy and sleep-deprived wild-type (WT-SD) mice. Veh, vehicle; HG, HG-9-91-01. d-g, Global (d, f) and quantitative (e, g) ΔPs analysis of HG/vehicle (Slp) and HG/ vehicle (WT-SD) groups. h-k, Circadian (h, j) and mean (i, k) absolute NREMS delta power analysis of HG/vehicle Sleepy (n = 14) and HG/ vehicle WT-SD (n = 16) groups. I, Molecular model of synaptic homeostasis and sleep-wake homeostasis. Mean, one-way ANOVA, Tukey's test (e, g); Mean \pm s.e.m., two-way ANOVA, Sidak's test (\mathbf{h}, \mathbf{j}) ; Paired t-test, two-tailed (\mathbf{i}, \mathbf{k}) .

*P < 0.05; †P < 0.01; ‡P < 0.001.

SD6/S6 and *Sleepy*/wild-type comparisons, respectively (Fig. 2a). Hyperphosphorylation of synapsin-1 in *Sleepy* and SD6 brain lysates was confirmed by phospho-tag gel electrophoresis (Fig. 2b).

Next, we performed global phosphorylation state change analysis for all quantified phosphoproteins in our datasets (Fig. 2c–e and Supplementary Table 3). In the sleep-deprived model, the phosphorylation state of 151 and 45 proteins is significantly upregulated (hyperphosphorylated, $\Delta Ps > 2.4$) in SD6 brains relative to RS3 or S6 brains, respectively (Fig. 2c, d). The phosphorylation state of 190 proteins is significantly upregulated, whereas the phosphorylation state of 52 proteins is downregulated (hypophosphorylation, $\Delta Ps < -2.4$) in the brains of *Sleepy* mice in comparison to those of wild-type mice (Fig. 2e). Cross-comparison of sleep-deprived and *Sleepy* models identified 80 hyperphosphorylated proteins, which we termed the sleep-need-index-phosphoproteins (SNIPPs), whose cumulative changes in phosphorylation state parallel those of sleep need in both models (Extended Data Fig. 7a).

Notably, 69 (>86%) of the 80 SNIPPs are annotated as synaptic proteins (Fig. 2f, Extended Data Fig. 7b and Supplementary Table 4a, b), whereas only 20% of the total phosphoproteins are annotated as synaptic proteins. A literature search reveals that mutations of 12 (15%) of the 80 SNIPPs cause sleep phenotypes in mice or humans (Fig. 2g and Supplementary Table 4a). Furthermore, we analysed published phosphoproteomic data of post-synaptic density (PSD) fractions from mouse forebrains collected in normal sleep (S4) and wake (W4) states⁴ (Fig. 2h and Supplementary Table 5). Approximately 70% of phosphorylation changes observed in PSD fractions are increases, and the mean Δ Ps value of the 80 SNIPPs is significantly increased in accordance with higher sleep need in wake brains relative to sleep brains (Fig. 2i and Extended Data Fig. 6c, d). These observations suggest a potential mechanistic link between the synaptic phosphoproteome and homeostatic sleep regulation (Supplementary Discussion 2).

Because synaptic activities underlie waking experience, we hypothesize that SNIPPs track waking experience through cumulative phosphorylation. To test this hypothesis, we conducted a time-course

sleep deprivation followed by quantitative phosphoproteomic analysis (Fig. 3a). Comparison of SD1, SD3 or SD6 and SD0 samples reveals a time-dependent increase in the number of phosphorylation events in whole-brain phosphoproteome (Fig. 3b). Δ Ps analysis indicates that the mean phosphorylation states of 80 SNIPPs gradually rise with the duration of sleep deprivation (Extended Data Fig. 6e), with many SNIPPs showing time-dependent cumulative phosphorylation (Fig. 3c, class A–C).

MK801, a specific inhibitor of *N*-methyl-D-aspartate receptor (NMDAR), has previously been identified as a potent inducer of SWA in rodents^{24–26}. Our quantitative phosphoproteomic analysis of this pharmacological model identified 31 hyperphosphorylated proteins (Δ Ps > 2.4) in the MK801 model compared to vehicle-only control, of which 25 (80%) are annotated as synaptic proteins (Fig. 3d, e and Extended Data Fig. 8). The MK801, *Sleepy* and sleep-deprived models have 21 SNIPPs in common (Extended Data Fig. 8j), 13 of which accumulate phosphorylation in a time-dependent manner (Fig. 3f). These 13 SWA-SNIPPs not only serve as a reliable molecular indicator of SWA or sleep need in multiple models, but also may contribute critically to regulation of SWA, a macro-electrophysiological readout of synaptic functions^{5,7} (Supplementary Discussion 3).

To examine whether SNIPPs are substrates of SLEEPY (the protein encoded by $Sik3^{Slp}$), we compared the interactomes of SLEEPY and wild-type SIK3 by immunoprecipitation and mass spectrometric analysis using whole-brain lysates from Flag-HA- $Sik3^{Slp}$ and Flag-HA- $Sik3^{+}$ knock-in mice¹² (Extended Data Fig. 9a and Supplementary Table 6). SLEEPY preferentially associated with synaptic proteins, including 28 of 80 SNIPPs (Fig. 4a and Extended Data Fig. 9b, c). Immunoprecipitation and western blotting confirmed enhanced associations between SLEEPY and SNIPPs such as the pre-synaptic active zone protein bassoon, synaptic RAS GTPase-activating protein 1 (SYNGAP1) and NMDAR subunits NR2B and NR1 (Fig. 4b).

We applied the AMPK Motif Analyzer to predict 2,943 phosphopeptides as potential AMPK substrates²⁷ in the *Sleepy*/wild-type phosphoproteome dataset (Extended Data Fig. 9d and Supplementary

Table 4c). Among these, 625 phosphopeptides were significantly changed (Q < 0.2) in Sleepy brains in comparison to wild-type brains, 462 of which were hyperphosphorylated in Sleepy brains (Extended Data Fig. 9d). The 28 SNIPPs that interact with SLEEPY contain 47 putative AMPK sites that are differently phosphorylated in Sleepy brains, of which 40 (85%) are hyperphosphorylated in Sleepy brains (Extended Data Fig. 9e). Recombinant SLEEPY and SIK3 exhibited similar in vitro kinase activities (Extended Data Fig. 9f), suggesting that SLEEPY itself does not have increased kinase activity. Taken together, these observations suggest that SLEEPY may increase phosphorylation of SNIPPs by enhancing kinase–substrate association.

Next, we attempted to rescue the phenotypes of *Sleepy* mice by intracerebroventricular injection of the pan-SIK inhibitor HG-9-91-01²⁸ (HG) to inhibit SLEEPY or SIK3 kinase activity (Fig. 4c). Administration of HG significantly reduced phosphorylation of AMPK substrates, particularly phosphorylation of the 28 SLEEPY-interacting SNIPPs (Extended data Fig. 9g–i). Consistent with this, HG treatment of *Sleepy* mice reduced phosphorylation of SNIPPs and SWA, but not duration, of NREMS (Fig. 4d, e, h, i and Extended Data Fig. 9j–m). Similarly, HG treatment of sleep-deprived wild-type mice reduced phosphorylation of AMPK substrates, phosphorylation of SNIPPs and SWA of NREMS (Fig. 4f, g, j, k and Extended Data Fig. 10), suggesting that SIK3 and SNIPPs have a critical role in normal homeostatic sleep regulation.

We hypothesize that a core set of SNIPPs monitor the duration and richness of prior waking through cumulative phosphorylation, which translates into a corresponding sleep need that determines the quality and duration of subsequent sleep¹¹ (Fig. 4l and Supplementary Discussion 4). Whereas prolonged wakefulness leads to cognitive impairment and sleepiness, sleep refreshes the brain through multiple restorative effects and optimizes cognitive functions for the next waking period^{5,7,9,11}. Specifically, the synaptic homeostasis hypothesis posits that waking encodes memories by potentiating synapses, whereas sleep consolidates memories and restores synaptic homeostasis by global downscaling of synaptic strength⁵. We hypothesize that the phosphorylation–dephosphorylation cycle of SNIPPs represent a major regulatory mechanism that underlies both synaptic homeostasis and sleep–wake homeostasis to maximize cognitive functions of the brain.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0218-8.

Received: 5 June 2017; Accepted: 1 May 2018; Published online 13 June 2018.

- Cirelli, C. & Tononi, G. Changes in anti-phosphoserine and antiphosphothreonine antibody binding during the sleep-waking cycle and after lesions of the locus coeruleus. Sleep Res. Online 1, 11–18 (1998).
- Elliott, A. S., Huber, J. D., O'Callaghan, J. P., Rosen, C. L. & Miller, D. B. A review of sleep deprivation studies evaluating the brain transcriptome. Springerplus 3, 728 (2014).
- Thompson, C. L. et al. Molecular and anatomical signatures of sleep deprivation in the mouse brain. Front. Neurosci. 4, 165 (2010).
- Diering, G. H. et al. Homer1a drives homeostatic scaling-down of excitatory synapses during sleep. Science 355, 511–515 (2017).
- Tononi, G. & Cirelli, C. Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron* 81, 12–34 (2014).
- de Vivo, L. et al. Ultrastructural evidence for synaptic scaling across the wake/ sleep cycle. Science 355, 507–510 (2017).
- Vyazovskiy, V. V. & Harris, K. D. Sleep and the single neuron: the role of global slow oscillations in individual cell rest. Nat. Rev. Neurosci. 14, 443–451 (2013).
- Borbely, A. A. A two process model of sleep regulation. Hum. Neurobiol. 1, 195–204 (1982).
- Benington, J. H. Sleep homeostasis and the function of sleep. Sleep 23, 959–966 (2000).

- Franken, P., Chollet, D. & Tafti, M. The homeostatic regulation of sleep need is under genetic control. J. Neurosci. 21, 2610–2621 (2001).
- Vassalli, A. & Dijk, D. J. Sleep function: current questions and new approaches. Eur. J. Neurosci. 29, 1830–1841 (2009).
- Funato, H. et al. Forward-genetics analysis of sleep in randomly mutagenized mice. Nature 539, 378–383 (2016).
- Saper, C. B. & Fuller, P. M. Wake-sleep circuitry: an overview. Curr. Opin. Neurobiol. 44, 186–192 (2017).
- Liu, S., Liu, Q., Tabuchi, M. & Wu, M. N. Sleep drive is encoded by neural plastic changes in a dedicated circuit. Cell 165, 1347–1360 (2016).
- Lizcano, J. M. et al. LKB1 is a master kinase that activates 13 kinases of the AMPK subfamily, including MARK/PAR-1. EMBO J. 23, 833–843 (2004).
- Erickson, B. K. et al. Evaluating multiplexed quantitative phosphopeptide analysis on a hybrid quadrupole mass filter/linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* 87, 1241–1249 (2015).
- McAlister, G. C. et al. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* 86, 7150–7158 (2014).
- Weekes, M. P. et al. Quantitative temporal viromics: an approach to investigate host-pathogen interaction. Cell 157, 1460–1472 (2014).
- Paulo, J. A. et al. Effects of MEK inhibitors GSK1120212 and PD0325901 in vivo using 10-plex quantitative proteomics and phosphoproteomics. *Proteomics* 15, 462–473 (2015).
- Humphrey, S. J., James, D. E. & Mann, M. Protein phosphorylation: a major switch mechanism for metabolic regulation. *Trends Endocrinol. Metab.* 26, 676–687 (2015).
- Greengard, P., Valtorta, F., Czernik, A. J. & Benfenati, F. Synaptic vesicle phosphoproteins and regulation of synaptic function. *Science* 259, 780–785 (1993).
- Česca, F., Baldelli, P., Valtorta, F. & Benfenati, F. The synapsins: key actors of synapse function and plasticity. *Prog. Neurobiol.* 91, 313–348 (2010).
- Cantrell, A. R. et al. Molecular mechanism of convergent regulation of brain Na⁺ channels by protein kinase C and protein kinase A anchored to AKAP-15. Mol. Cell. Neurosci. 21, 63–80 (2002).
- Tatsuki, F. et al. Involvement of Ca²⁺-dependent hyperpolarization in sleep duration in mammals. *Neuron* 90, 70–85 (2016).
- Campbell, I. G. & Feinberg, I. NREM delta stimulation following MK-801 is a response of sleep systems. J. Neurophysiol. 76, 3714–3720 (1996).
- Campbell, I. G. & Feinberg, I. Noncompetitive NMDA channel blockade during waking intensely stimulates NREM delta. J. Pharmacol. Exp. Ther. 276, 737–742 (1996)
- Schaffer, B. E. et al. Identification of AMPK phosphorylation sites reveals a network of proteins involved in cell invasion and facilitates large-scale substrate prediction. Cell Metab. 22, 907–921 (2015).
- Clark, K. et al. Phosphorylation of CRTC3 by the salt-inducible kinases controls the interconversion of classically activated and regulatory macrophages. Proc. Natl Acad. Sci. USA 109, 16986–16991 (2012).

Acknowledgements We are grateful to M. Dong, S. Chen and H. Mirzaei for mass spectrometry assistance; J. Cohen, R. Greene and F. Shao for comments on the manuscript. Q.L. is a W.A. 'Tex' Moncrief Jr. Scholar in Medical Research. Y.Y. is a Virginia Murchison Linthicum Scholar in Medical Research and a CPRIT scholar in Cancer Research. This work was supported by the Welch foundation (I-1608 to Q.L.; I-1800 to Y.Y.), the National Institute of Health (GM111367 to Q.L.; R01AG047928 to J.P.; GM114160 to Y.Y.), JSPS KAKENHI (16K16639 to Z.W.; 17K15592 to J.M.; 26220207, 17H06095 to M.Y., H.F., Q.L.; 17H04023, 16K15187, 15H05942 to H.F.), JST CREST (JPMJCR1655 to M.Y.), FIRST program from JSPS to M.Y., Uehara and Takeda Foundations to M.Y. and the WPI program from Japan's MEXT.

Reviewer information *Nature* thanks D. Kirkpatrick and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Z.W., J.M. and Q.L. designed experiments with inputs from M.Y., H.F. and L.C. Z.W. received mass spectrometric training from Y.Y. Z.W., Y.L. and C.L. performed bioinformatics analysis with advice from Y.O. and J.P. J.M., Z.W., C.Ma and X.Y. performed biochemical studies. C.Mi., M.K., A.I., N.H.-H., S.K., X.G., J.M., Z.W. collected tissue samples for mass spectrometry. J.M., Z.W., T.L., X.G., S.Z. and M.S. completed EEG/EMG data analysis. D.K.-N., T.F. and S.T. produced genetically modified mice. J.M. and Z.W. made the figures. Q.L. and Z.W. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0218-8.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0218-8.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to H.F. or M.Y. or O.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

General materials. Tandem mass tag (TMT) isobaric reagents, water and organic solvents were purchased from Thermo Fisher Scientific. Titansphere titanium dioxide (TiO₂) beads were from GL Sciences. Phospho-tag was from Wako Pure Chemical Industries. Unless otherwise noted, all other chemicals were from Sigma-Aldrich or Nacalai Tesque.

Animal studies. All animal experiments were performed according to procedures approved by the Institutional Animal Care and Use Committee of University of Tsukuba or University of Texas Southwestern Medical Center at Dallas. All mice used were males on a C57BL/6N background and were housed under humidity-and temperature-controlled conditions (22–25 \pm 1 °C) on a 12-h light–dark cycle. Food and water were provided ad libitum.

Sleep phenotype analysis. The sleep-wake behaviours were analysed as previously described with modifications¹². Electroencephalogram (EEG)/electromyogram (EMG) data were visualized and analysed using a custom semi-automated staging MatLab (MathWorks)-based program, followed by visual inspection. We did not apply blinding and only excluded animals with unreadable EEG signals from final sleep analysis. In brief, mice were implanted with the EEG/EMG electrodes at the age of 8-10 weeks, and EEG/EMG signals were recorded during weeks 12-20. Age-matched control and treatment groups of animals were used for each experiment. Following semi-automated analysis of EEG/EMG data, EEG signals were subjected to fast Fourier transform analysis for 1 to 30 Hz with 1-Hz bins. Wake was defined by low amplitude, fast EEG and high amplitude, variable EMG; NREMS by high amplitude, delta (1-4Hz) frequency EEG and low EMG tonus; and REMS by dominant theta (6-9 Hz) frequency EEG and EMG atonia. Absolute and relative power spectrum analyses of corresponding states within indicated ZT times were performed; for relative power spectrum analysis (%), the EEG power of each frequency bin was expressed as a percentage of the total power over all frequency bins (1-30 Hz). Absolute NREMS delta power density (arbitrary units) is determined by the delta band of NREMS and normalized to the average NREMS delta power during ZT8 to ZT11 of the baseline recording day¹⁰; relative delta power density (%) is defined by the ratio of delta power (1-4 Hz) to total power of NREMS EEG. In circadian variation plots, each data point represents the mean value of NREMS delta power or duration in the following 1 h.

Experimental design. To examine how different treatments affect sleep/wake behaviours, 3-day baseline EEG/EMG recordings were conducted after mice were acclimated for a week. Mice remained in the same recording chamber for a 3-6-day interval between treatments. No abnormal EEG/EMG signals were confirmed during the interval before next treatment.

For the sleep deprivation model, mice were sleep deprived on an automated orbital shaker with access to food and water 12. A 1-day baseline recording taken before sleep deprivation was used as the basal condition. Whole brains or livers were collected at ZT6 for ad libitum sleep (S6) and sleep-deprived (SD6, ZT0–ZT6) wild-type mice, or at ZT9 for 6-h sleep deprivation followed by 3-h recovery sleep (RS3) wild-type mice. For time-course sleep-deprivation, whole brains of wild-type mice were collected at ZT0 (SD0) or after 1, 3 or 6h of sleep deprivation (SD1, SD3, SD6). For the Sleepy model, baseline EEG/EMG recording data were used; whole brains were collected for Sik3^{+/+} (wild-type) and Sik3^{Slp/+} (Sleepy) at ZT12.5. For food/water deprivation experiments, sham deprived (old food/water exchanged for new food/water) and deprived (all food/water removed at indicated ZT) were conducted in both normal sleep and sleep-deprived conditions; whole brains were collected at ZT6 for both conditions. For MK801 treatment, we performed intraperitoneal injection of mice with vehicle (0.9% saline) followed by 2 mg/kg MK801 (Sigma-Aldrich). Wild-type mice were injected at ZT17 in the previous dark phase followed by EEG/EMG recording at the onset of light phase (ZT0); whole brains were collected at ZT23.5, 6.5 h after MK801 administration. For HG-9-91-01²⁸ (ApexBio) treatment, we performed intracerebroventricular injection of mice with vehicle (3% DMSO) followed by 8 mg/kg HG-9-91-01. Sik3^{Slp/+} mice were injected at ZT6 and ZT9; whole brains were collected at ZT11.5. Wild-type mice were injected at ZT0 and ZT3 during sleep deprivation (ZT0-ZT6); whole brains were collected at ZT6. The organization of sleep experiments and sleep phenotype results are listed in Supplementary Table 7b.

Tissue lysate preparation. Mouse tissues (whole brain or liver) were quickly dissected at indicated ZT, rinsed with PBS and flash frozen in liquid nitrogen. Typically, one mouse brain was homogenized in a glass tissue homogenizer with 5 ml of lysis buffer (50 mM HEPES, pH 7.4, 150 mM NaCl, 2.5% SDS, 2 mM MgCl₂) freshly supplemented with protease and phosphatase inhibitor cocktail tablets (Roche). Tissue homogenates were incubated at room temperature for 30 min and centrifuged at 15,000g for 20 min. The supernatant was carefully transferred to a new tube without disturbing the pellet. Protein concentration of protein lysates was determined using the bicinchoninic acid (BCA) assay (Thermo Scientific Pierce).

For comparison of the SIK3 and SLEEPY interactomes, wild-type, Flag-HA- $Sik3^+$ and Flag-HA- $Sik3^{Slp}$ knock-in mouse brains were lysed in ice-cold lysis buffer (20 mM HEPES, pH 7.4, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100,

 $2\,\text{mM}$ MgCl₂, $15\,\text{mM}$ NaF, $10\,\text{mM}$ Na₄P₂O₇) freshly supplemented with protease/phosphatase inhibitors in a glass tissue homogenizer 12 . After $30\,\text{min}$ incubation on ice, brain homogenates were centrifuged at 13,000g for $20\,\text{min}$ at $4\,^{\circ}\text{C}$. The supernatant was pre-cleared with IgG and Protein G beads for $30\,\text{min}$ before immunoprecipitation. $50\,\mu\text{l}$ of anti-Flag antibody-conjugated Sepharose beads (A2220, Sigma-Aldrich) was added to each pre-cleared lysate and rotated overnight at $4\,^{\circ}\text{C}$. The beads were washed five times with cold wash buffer (20 mM HEPES, pH 7.4, $150\,\text{mM}$ NaCl, $1\,\text{mM}$ EDTA, 1% Triton X-100, $2\,\text{mM}$ MgCl₂, $15\,\text{mM}$ NaF, $10\,\text{mM}$ Na₄P₂O₇), $50\,\mu\text{l}$ of elution buffer (2% SDS, $60\,\text{mM}$ Tris-HCl, pH 6.8, $50\,\text{mM}$ dithiothreitol (DTT), 10% glycerol) was then added and rotated for $10\,\text{min}$ at $^{\circ}\text{C}$. Protein elution was repeated twice and combined into one eluate, then analysed by mass spectrometry and western blotting.

Mass spectrometry sample preparation. Protein lysate sample was reduced with DTT and then alkylated with iodoacetamide. Chloroform—methanol precipitation of protein lysate was performed, and the precipitate was resuspended in 8 M urea buffer. Protein lysate was digested for 2 h with Lys-C (1:100, enzyme to protein; Wako), followed by dilution to 2 M urea with 25 mM ammonium carbonate buffer (pH 7.8), and trypsin (1:100, enzyme to protein; Thermo Scientific Pierce) digestion overnight at room temperature. After stopping the digestion with 1% formic acid, the peptide mixture was subjected to C18 solid-phase extraction (Sep-Pak, Waters) for desalting, and subsequently vacuum-centrifuged to near-dryness.

For phosphopeptide enrichment, desalted peptides were resuspended in 1 ml phosphopeptide binding buffer (2 M lactic acid/50% acetonitrile (ACN)) and centrifuged at 15,000g for 20 min at room temperature. The supernatant was carefully transferred to a new tube without disturbing the pellet. TiO $_2$ beads were washed three times with phosphopeptide binding buffer, added to the supernatant (peptide mixture) and incubated with gentle rotation for 1 h at room temperature. Afterwards, TiO $_2$ beads were washed twice with phosphopeptide binding buffer and twice with wash buffer (50% ACN/0.1% trifluoroacetic acid). Phosphopeptides were eluted twice from TiO $_2$ beads with 500 μ l elution buffer (50 mM $\rm K_2HPO_4$, pH 10), acidified with 20% formic acid, subjected to desalting and vacuum-centrifuged to near-dryness.

Desalted peptides were resuspended in 200 mM HEPES (pH 8.5) and peptide concentration was determined using the BCA assay. Approximately $50\,\mu g$ of peptides for each sample were labelled with TMT reagent for 1 h at room temperature. After the reaction was quenched with hydroxylamine, all TMT-labelled samples for one experiment were combined into one mixture, acidified with 20% formic acid, desalted and vacuum-centrifuged to near-dryness. The TMT-labelled sample mixture was solubilized in HPLC buffer A (1% ACN, 10 mM ammonium bicarbonate, pH 8.0) for HPLC fractionation using an Agilent 300 Extend C18 column (5- μ m particles, 4.6-mm internal diameter, 150 mm in length). Different HPLC fractions were acidified with 20% formic acid and vacuum centrifuged to near-dryness. Each fraction was desalted using a StageTip, dried by vacuum centrifugation and re-suspended for LC/MS analysis.

Mass spectrometry data acquisition. Data were collected using the Orbitrap-Fusion mass spectrometry platform coupled with EASY-nLC 1000 liquid chromatography pump (Thermo Fisher Scientific). A pre-column (Acclaim PepMap 100 C18, Thermo Fisher Scientific) and analytical column (NTCC-360/75-3-125, NIKKYO) were used for sample trapping and analytical separation. Peptides were separated at a flow rate of 300 nl/min using a gradient of 6–27% ACN (0.1% formic acid) over 190 min.

The MultiNotch synchronous precursor selection MS³-based TMT method was used on an Orbitrap-Fusion mass spectrometer using Xcalibur (v.3.0; Thermo Fisher Scientific) as described with modifications $^{16-19}$. In brief, first stage of mass spectrum data between 400–1500 m/z were acquired from the Orbitrap at 120,000 resolution in profile data type with 4e5 AGC target, 50-ms maximum injection time. Ions were isolated in top speed mode using the quadrupole with a 0.7-m/z isolation window. MS² scans between 400–1200 m/z were acquired from the ion trap in centroid data type with CID fragmentation (35% collision energy) in Turbo mode, 1e4 AGC target, 50-ms maximum injection time. Top ten MS² fragment ions were selected using synchronous precursor selection mode for TMT reporter ions quantitation. MS³ scans were acquired from the Orbitrap at 60,000 resolution in profile data type with HCD fragmentation (65% collision energy), 1e5 AGC target, 120-ms maximum injection time. Ions were not accumulated for all parallelisable time.

Mass spectrometry data analysis. Raw mass spectrometry files from the entire study were searched against a composite target/decoy database using SEQUEST²⁹⁻³¹ from Proteome Discoverer software (PD, v.2.1, Thermo Fisher Scientific). The target mouse protein database was generated from UniProt, combining all Swiss-Prot and TrEMBL entries (17 October 2015). $\rm MS^2$ spectra were searched with \pm 20 ppm for precursor ion mass tolerance, \pm 1 Da for fragment ion mass tolerance, fully tryptic restriction, four maximal missed cleavages, dynamic mass shift for oxidation of methionine (+15.9949 Da), fixed TMT modifications on the N terminus and lysine (+229.1629 Da), and carbamidomethylation of

cysteine residues (+57.0215 Da). For phosphoproteomic analysis, additional dynamic modifications on serine, threonine and tyrosine (+79.9663 Da) were used. The peptide spectrum matches (PSMs) were filtered by Percolator 32 (PD 2.1) to achieve 1% protein and peptide FDR (according to Q value) for proteome and phosphoproteome, respectively. ptmRS 33 (PD 2.1) was used for phosphorylation site localization, which derived a localization probability score for each putatively modified site based on the given MS^2 data. Phosphopeptides with phosphorylation site probability score \geq 25 were considered in following analysis.

TMT reporter ion signal-to-noise (S/N) values were quantified from MS^3 scans using an integration tolerance of 20 ppm (Orbitrap) with the most confident centroid setting (PD 2.1) for matching peptides. For interactome analysis, raw reporter ion abundance was used for further analysis. For proteomic and phosphoproteomic analysis, the sum of raw reporter ion for each channel was normalized assuming equal input loading of all channels. The sum of reporter ions for each protein was used in protein quantitation. The normalized quantification data of all quantified proteins, peptides or phosphopeptides were used for further analysis.

To evaluate the confidence of protein identification and quantification by PD, we used a recently developed proteomics pipeline JUMP^{34,35} (v.1.12.1) to re-process one set of proteome data (EX4, SlpWTpa2) with the above same database search and PSM filtering parameters. The consistency of protein quantification between these two pipelines was indicated by Pearson correlation, which was calculated for each PSM from proteins quantified by both pipelines.

Proteomic data processing. For proteomic analysis, different isoforms were considered as different proteins for data analysis unless otherwise stated. For phosphoproteomic analysis, phosphopeptide was used for further analysis including unique and composite (containing \geq 2 phosphorylation sites) forms. The normalized quantification data of all quantified proteins, peptides or phosphopeptides were consolidated (sum of value) to generate a unique subject ID. The consolidated abundance values were then scaled for each protein or phosphopeptide so that the average abundance was one. The scaled data from different TMT-multiplex experiments for the same comparison (for example, Sleepy/wild-type group) were integrated together based on unique subject ID. The multiple unpaired t-test (P value) analysis followed by the two-stage step-up FDR (Q value) approach was used to determine statistical significance (Q < 0.2) for each comparison³⁶. The mean value for each experiment condition was used to generate the log₂(fold change) value for each unique subject, which was used for further analysis. To evaluate phosphorylation stoichiometry³⁷, phosphoproteome normalization was performed for SD6/RS3, SD6/S6 and Sleepy/wild-type groups for which whole proteome and phosphoproteome data were available. In brief, the scaled phosphopeptide abundances of SD6 and Slp groups were adjusted with the mean abundance fold-change value of corresponding protein. Pearson correlation of log₂(fold change) value between normalized and un-normalized was performed to evaluate the normalization effect. The full description and datasets for all proteomic experiments are listed in Supplementary Table 1, and those for all phosphoproteomics experiments are listed in Supplementary Table 2.

For SIK3 and SLEEPY interactome analysis, raw abundance data of all quantified proteins were consolidated (sum of value) to generate unique subject IDs, and then normalized assuming equal SIK3/SLEEPY protein amount in all channels. Two criteria were used to define the interacting protein (ip) for SIK3 or SLEEPY: a) TMT intensity [Mean—Blank > 10]; b) fold change [Mean/Blank > 2]. For the SIK3 preferential interacting protein (SIK3-pip): (Mean SLEEPY—Blank)/(Mean SIK3—Blank) < 0.5; SLEEPY-pip: (Mean SLEEPY—Blank)/(Mean SIK3—Blank) > 2. SLEEPY-pip proteins were used for the Gene Ontology (GO) cellular component enrichment analysis through Gene Ontology Consortium and PANTHER classification system $^{38-40}$. All 22,262 genes of $Mus\ musculus$ in the database were used as reference to determine the fold enrichment. Fisher's exact with FDR multiple test correction was used to determine statistical significance. The full description and datasets are listed in Supplementary Table 6.

Protein phosphorylation-state analysis. The phosphorylation state change (ΔPs) value for individual proteins is calculated as the sum of log2(fold change) value of all phosphopeptides with statistically significant changes (Q < 0.2) from all protein isoforms encoded by the same gene. If none of a phosphopeptide's Q values is above 0.2, the ΔPs value will be zero. The total quantified phosphopetides number of the SD6/RS3 group was used for the ΔPs value normalization with other comparisons of brain phosphoproteome in this study. Normalized ΔPs value was used for further analysis to determine the hyperphosphorylated or hypophosphorylated proteins. To set up the cutoff for ΔPs value, two null tests were performed using data from SD6/RS3 and Sleepy/wild-type phosphoproteomes, briefly, data from channels with even number between two groups were swapped to determine the FDR. For two null tests, no phosphopeptide has a Q value above 0.2 and Δ Ps value is zero for all proteins (Supplementary Table 2z, aa). Because average standard deviation (s.d.) for ΔPs value of SD6/RS3, SD6/S6 and Sleepy/wild-type groups is 1.1 (Supplementary Table 3d), we applied a stringent cut-off for ΔPs value at ± 2.4 (>2 s.d.) for each comparison group to represent the concept of cumulative phosphorylation. Hyperphosphorylated (hyper, $\Delta Ps > 2.4$), hypophosphorylated (hypo, $\Delta Ps < -2.4$) phosphoproteins. The full description and datasets are listed in Supplementary Table 3.

As previously described⁴, in the normal sleep–wake model, mouse forebrains (cortex plus hippocampus) were collected at ZT16 (W4) and ZT4 (S4) to purify PSD fractions for phosphoproteomic analysis. For analysis of the normal sleep–wake model, the raw phosphopeptide data from supplementary tables 2A (hyperphosphorylated during wake (10pm/10am ratio >1.3)) and S2B (hyperphosphorylated in the PSD during sleep (10am/10pm ratio >1.3, or 10pm/10am ratio <0.77)) in ref. 4 were combined into one data table. The raw quantification data of all phosphopeptides were consolidated (sum of value) to generate a unique subject ID and log₂ (fold change) value. It should be noted that no statistical test was performed for phosphopeptide comparisons as there were only two technical replicates for each condition. The ΔPs value for each protein is calculated as the sum of log₂(fold change) value of all phosphopeptides from all protein isoforms encoded by the same gene, which was not normalized with SD6/RS3 group. The full description and datasets were listed in Supplementary Table 5.

Bioinformatics analysis. The sleep phenotypes, molecular and neuronal functions of 80 SNIPPs were classified manually by literature mining $^{23,24,41-55}$, the complete literature information is listed in Supplementary Table 4a. The classification of synaptic proteins was mainly based on an integrated synaptic protein database from 11 proteomics studies $^{56-66}$ as listed in Supplementary Table 4b. A protein that is shown in ≥ 2 references (Synaptic Ref Count ≥ 2) is considered as an annotated synaptic protein. To predict potential AMPK substrates, a sequence window of -5 to +4 positions around each phosphorylation site was scored with the AMPK motif analyzer 27 . Putative AMPK substrate) were used for further analysis. Complete data for AMPK substrate prediction was listed in Supplementary Table 4c. Hierarchical clustering (centroid linkage with Euclidean distance) was performed with Cluster 3.0^{67} .

In vitro kinase assay. The kinase activities of recombinant SIK3 and SLEEPY proteins were measured by in vitro kinase assay as previously described 68 . A recombinant GST–MFF(S146) (136-RQNGQLVRNDSIVTPSPPQA-155; AMPK motif score = 1.06) fusion protein was used as substrate. Recombinant Flag-SIK3 and Flag-SLEEPY were overexpressed in HEK 293T cells and affinity purified with anti-Flag antibody-conjugated Sepharose beads. A mock preparation from HEK293T cells transfected with empty vector was used as negative control. The same amount of recombinant kinase and substrate proteins were incubated for 20 min at 30 °C in kinase reaction buffer (50 mM HEPES, pH 7.4, 1 mM EDTA, 10 mM MgCl₂, 0.5 mM ATP) freshly supplemented with protease/phosphatase inhibitors. Reactions were stopped by the addition of sample loading buffer; samples were resolved by SDS–PAGE followed by western blotting or by Coomassie blue staining.

Phospho-tag SDS-PAGE and immunoblotting. Equal amounts of protein samples were resolved by phospho-tag⁶⁹ (Wako) or SDS-PAGE and transferred to PVDF membrane. Phos-tag SDS-PAGE is an electrophoresis technique capable of separating phosphorylated and non-phosphorylated forms based on phosphorylation levels, owing to binding to the phospho-tag chemical, which slows the migration of phosphorylated protein⁶⁹. The molecular weight markers are only indicative for the non-phosphorylated forms and irrelevant for the phosphorylated forms. The Rf value of 1.0 is defined as the position of bromphenol blue dye⁶⁹.

Western blotting was performed according to standard procedures using the corresponding antibodies. Antibodies were used at the optimal concentration according to the manufacturer's instructions. Lamin A/C was measured as a loading control for the quantitative analysis of immunoblots of phosphorylation-motif antibodies. Antibodies used in this study included anti-EF2 (phospho T56/ T58) (ab82981, Abcam), anti-EF2 (#2332, Cell Signaling), anti-CaMKII (phospho T286) (ab32678, Abcam), anti-CaMKII (#4436, Cell Signaling), anti-nNOS (phospho S1417) (ab5583, Abcam), anti-nNOS (ab76067, Abcam), anti-KCC2 (phospho S940) (612-401-E15, Rockland), anti-KCC2 (07-432, EMD Millipore), anti-synapsin-1 (phospho S605) (#88246, Cell Signaling), anti-synapsin-1 (sc-8295, Santa Cruz), anti-phospho-AMPK Substrate Motif (LXRXX(S*/T*)) (#5759, Cell Signaling), anti-phospho-PKC substrate motif ((K/R)XS*X(K/R)) (#6967, Cell Signaling), anti-phospho-PKA substrate motif ((K/R)(K/R)X(S*/T*))(#9624, Cell Signaling), anti-phospho-ATM/ATR substrate motif (S*Q) (#9607, Cell Signaling), anti-phospho-AKT substrate motif (RXX(S*/T*)) (#9614, Cell Signaling), anti-phospho-PDK1 docking motif $((F/K)XX(F/Y)(S^*/T^*)(F/Y))$ (#9634, Cell Signaling), anti-phospho-CK2 substrate motif ((S*/T*)DXE) (#8738, Cell Signaling), anti-phospho-MAPK/CDK substrate motif (PXS*P, S*PX(K/R)) (#2325, Cell Signaling), anti-phospho-CDKs substrate motif ((K/H)S*P) (#9477, Cell Signaling), anti-phospho-PLK binding motif (ST*P) (#5243, Cell Signaling), anti-phospho-Thr-Pro motif (T*P, T*PP) (#3003, Cell Signaling), anti-phospho-Thr-Pro-Glu motif (T*PE, T*P) (#3004, Cell Signaling), anti-phospho-Thr-X-Arg motif (T*X(K/R)) (#2351, Cell Signaling), anti-phospho-Tyr (Y*) (#8954,



Cell Signaling), anti-Lamin A/C (sc-6215, Santa Cruz), anti-HA (Y-11) (sc-805, Santa Cruz), anti-NMDAR1 (MAB363, EMD Millipore), anti-NMDAR2B (75–101, NeuroMab), anti-SynGAP (#5539, Cell Signaling) and anti-SIK3 C-term, a custom-generated rabbit polyclonal antibody against the C-terminal 171 amino acids of mouse SIK3.

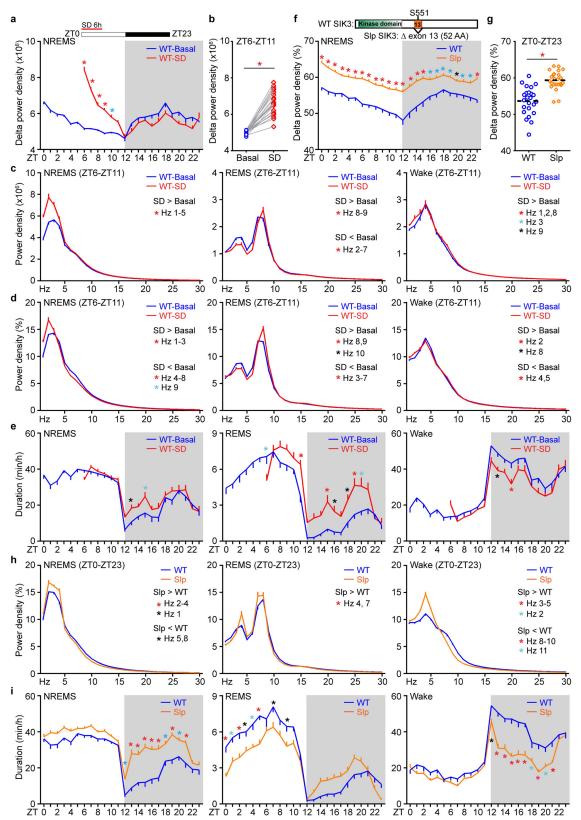
Statistical methods. Unless otherwise noted, all experimental subjects are biological replicates and at least two independent experiments were performed. ImageJ software was used to quantify intensity of protein bands. GraphPad Prism 7 or R software was used for statistical tests. No statistical methods were used to predetermine sample size. Randomization and blinding were not used. Following one-way or two-way analysis of variance (ANOVA), Fisher's LSD test compares one mean with another mean; Tukey's test compares every mean with every other mean; Dunnett's test compares every mean to a control mean; Sidak's test compares a set of means. Repeated measures or paired test was performed for matched subject comparisons. P < 0.05 was considered statistically significant. The complete sample size, statistical test method and results for each comparison are reported in the figure legends and described in detail in Supplementary Table 7a.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The mass spectrometry datasets, including raw data files, search engine files, full experimental summary file and Supplementary Tables 1 and 2, have been deposited to MassIVE^{70,71} with accession code MSV000081865 and to Proteome Xchange with accession code PXD008558. Source Data are provided with the online version of the paper. All other datasets generated and/or analysed in the current study are available from the corresponding author on reasonable request.

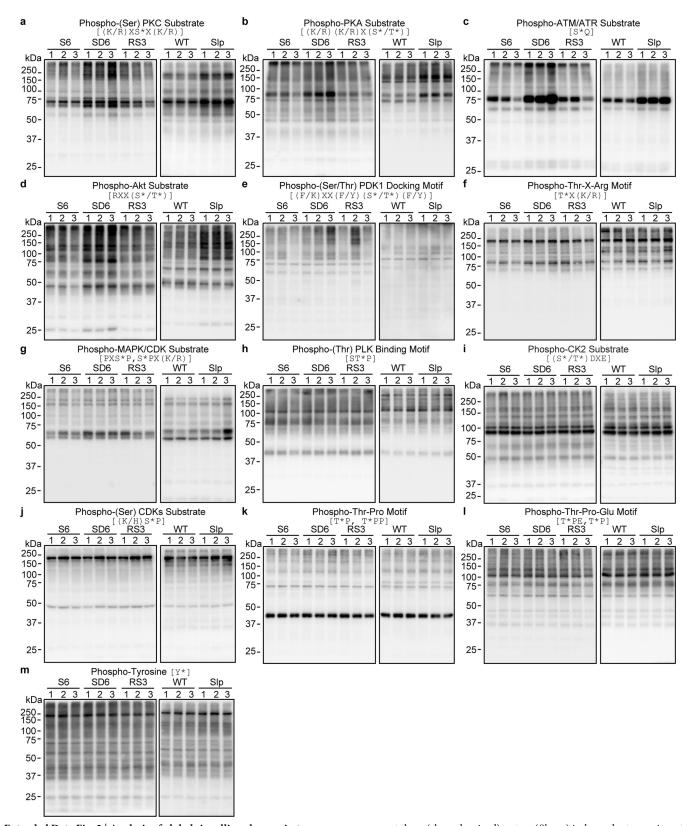
- Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989 (1994).
- Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC–MS/MS) for large-scale protein analysis: the yeast proteome. J. Proteome Res. 2, 43–50 (2003).
- Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 4, 207–214 (2007).
- Kall, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semisupervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* 4, 923–925 (2007).
- Taus, T. et al. Universal and confident phosphorylation site localization using phosphoRS. J. Proteome Res. 10, 5354–5362 (2011).
- Wang, X. et al. JUMP: a tag-based database search tool for peptide identification with high sensitivity and accuracy. Mol. Cell. Proteomics 13, 3663–3673 (2014).
- Li, Y. et al. JUMPg: an integrative proteogenomics pipeline identifying unannotated proteins in human brain and cancer cells. J. Proteome Res. 15, 2309–2320 (2016).
- Benjamini, Y., Krieger, A. M. & Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 93, 491–507 (2006).
- Wu, R. et al. Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes. *Mol. Cell. Proteomics* 10, M111 009654 (2011).
- 38. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
- The Gene Ontology Consortium. C. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 45, D331–D338 (2017).
- Mi, H. et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 45, D183–D189 (2017).
- Beacham, D., Ahn, M., Catterall, W. A. & Scheuer, T. Sites and molecular mechanisms of modulation of Na(v)1.2 channels by Fyn tyrosine kinase. J. Neurosci. 27, 11543–11551 (2007).
- James, T. F. et al. The Nav1.2 channel is regulated by GSK3. Biochim. Biophys. Acta 1850, 832–844 (2015).
- Siwek, M. E. et al. The CaV2.3 R-type voltage-gated Ca²⁺ channel in mouse sleep architecture. Sleep 37, 881–892 (2014).
- Parker, M. J. et al. De novo, heterozygous, loss-of-function mutations in SYNGAP1 cause a syndromic form of intellectual disability. Am. J. Med. Genet. A. 167A, 2231–2237 (2015).

- Carlisle, H. J. et al. Deletion of densin-180 results in abnormal behaviors associated with mental illness and reduces mGluR5 and DISC1 in the postsynaptic density fraction. *J. Neurosci.* 31, 16194–16207 (2011).
- Soorya, L. et al. Prospective investigation of autism and genotype–phenotype correlations in 22q13 deletion syndrome and SHANK3 deficiency. *Mol. Autism* 4, 18 (2013).
- Ahnaou, A., Raeymaekers, L., Steckler, T. & Drinkenbrug, W. H. Relevance of the metabotropic glutamate receptor (mGluR5) in the regulation of NREM–REM sleep cycle and homeostasis: evidence from mGluR5^{-/-}mice. *Behav. Brain Res.* 282, 218–226 (2015).
- Hagebeuk, E. E., van den Bossche, R. A. & de Weerd, A. W. Respiratory and sleep disorders in female children with atypical Rett syndrome caused by mutations in the CDKL5 gene. *Dev. Med. Child Neurol.* 55, 480–484 (2012).
- Lonart, G., Tang, X., Simsek-Duran, F., Machida, M. & Sanford, L. D. The role of active zone protein Rab3 interacting molecule 1 alpha in the regulation of norepinephrine release, response to novelty, and sleep. *Neuroscience* 154, 821–831 (2008).
- Iqbal, Z. et al. Homozygous and heterozygous disruptions of ANK3: at the crossroads of neurodevelopmental and psychiatric disorders. *Hum. Mol. Genet.* 22, 1960–1970 (2013).
- von Stulpnagel, C. et al. SYNGAP1 mutation in focal and generalized epilepsy: a literature overview and a case report with special aspects of the EEG. Neuropediatrics 46, 287–291 (2015).
- Mangatt, M. et al. Prevalence and onset of comorbidities in the CDKL5 disorder differ from Rett syndrome. Orphanet J. Rare Dis. 11, 39 (2016).
- Fehr, S. et al. The CDKL5 disorder is an independent clinical entity associated with early-onset encephalopathy. Eur. J. Hum. Genet. 21, 266–273 (2013).
- Jiang, P. et al. A systems approach identifies networks and genes linking sleep and stress: implications for neuropsychiatric disorders. *Cell Reports* 11, 835–848 (2015).
- Welch, J. M. et al. Cortico-striatal synaptic defects and OCD-like behaviours in Sapap3-mutant mice. Nature 448, 894–900 (2007).
- Bayes, A. et al. Comparative study of human and mouse postsynaptic proteomes finds high compositional conservation and abundance differences for key synaptic proteins. *PLoS ONE* 7, e46683 (2012).
- Li, J. et al. Long-term potentiation modulates synaptic phosphorylation networks and reshapes the structure of the postsynaptic interactome. Sci. Signal. 9, rs8 (2016).
- Uezu, A. et al. Identification of an elaborate complex mediating postsynaptic inhibition. *Science* 353, 1123–1129 (2016).
- Gonzalez-Lozano, M. A. et al. Dynamics of the mouse brain cortical synaptic proteome during postnatal brain development. Sci. Rep. 6, 35456 (2016).
- Weingarten, J. et al. The proteome of the presynaptic active zone from mouse brain. Mol. Cell. Neurosci. 59, 106–118 (2014).
- Boyken, J. et al. Molecular profiling of synaptic vesicle docking sites reveals novel proteins but few differences between glutamatergic and GABAergic synapses. *Neuron* 78, 285–297 (2013).
- Abul-Husn, N. S. et al. Systems approach to explore components and interactions in the presynapse. *Proteomics* 9, 3303–3315 (2009).
- Biesemann, C. et al. Proteomic screening of glutamatergic mouse brain synaptosomes isolated by fluorescence activated sorting. EMBO J. 33, 157–170 (2014).
- Distler, U. et al. In-depth protein profiling of the postsynaptic density from mouse hippocampus using data-independent acquisition proteomics. *Proteomics* 14, 2607–2613 (2014).
- Loh, K. H. et al. Proteomic analysis of unbounded cellular compartments: synaptic clefts. Cell 166, 1295-1307 (2016).
- 66. Nakamura, Y. et al. Proteomic characterization of inhibitory synapses using a novel pHluorin-tagged γ -aminobutyric acid receptor, type A (GABA_A), α 2 subunit knock-in mouse. *J. Biol. Chem.* **291**, 12394–12407 (2016).
- 67. de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453–1454 (2004).
- Lee, E. E. et al. A protein kinase C phosphorylation motif in GLUT1 affects glucose transport and is mutated in GLUT1 deficiency syndrome. Mol. Cell 58, 845–853 (2015).
- Kinoshita, E., Kinoshita-Kikuta, E., Takiyama, K. & Koike, T. Phosphate-binding tag, a new tool to visualize phosphorylated proteins. *Mol. Cell. Proteomics* 5, 749–757 (2006).
- 70. Vizcaino, J. A. et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014)
- Deutsch, E. W. et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* 45, D1100–D1106 (2017).



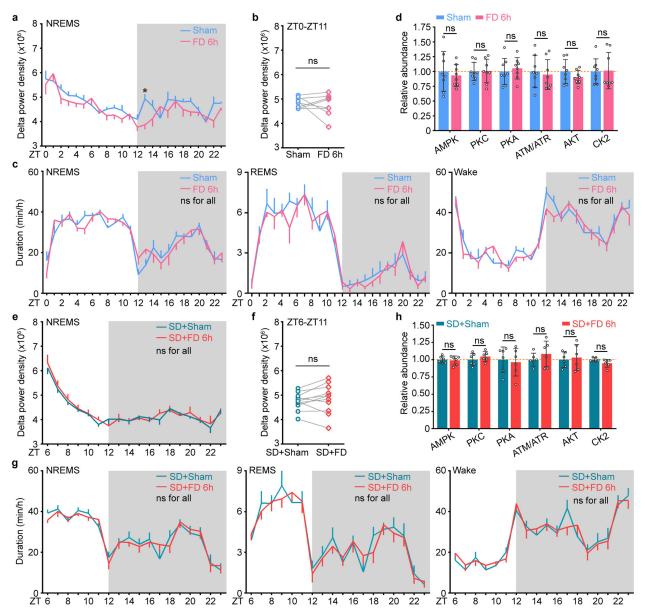
Extended Data Fig. 1 | Sleep phenotype analysis of the sleep-deprived and *Sleepy* models. a-e, Analysis of circadian (a) and mean (b) absolute NREMS delta power, absolute EEG power spectra (c), relative EEG power spectra (d) and duration (e) of NREMS, REMS and wake states of wild-type mice (n=24) without (WT-basal) and with 6 h of sleep deprivation (WT-SD). f-i, Analysis of circadian (f) and mean (g) relative

NREMS delta power, relative EEG power spectra (**h**) and duration (**i**) of NREMS, REMS and wake states of $Sik3^{+/+}$ (WT, n=24) and $Sik3^{Slp/+}$ (Slp, n=24) mice. Mean \pm s.e.m., two-way ANOVA with Sidak's test (**a**, **c**-**f**, **h**, **i**); Paired t-test, two-tailed (**b**); Mean, unpaired t-test, two-tailed (**g**). *(black), P < 0.05; *(cyan), P < 0.01; *(red), P < 0.001.



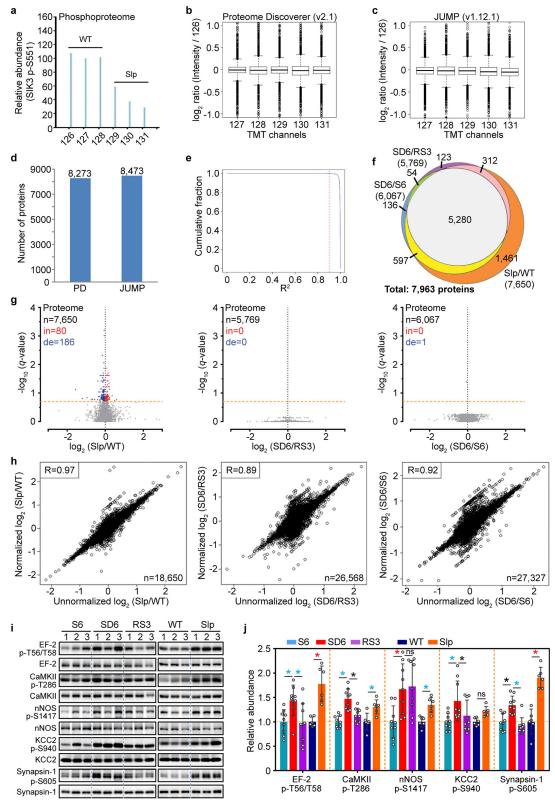
Extended Data Fig. 2 | Analysis of global signalling changes in two models of increased sleep need. a-m, Representative immunoblots using antibodies specific for 13 phosphorylation motifs to assess global signalling changes in whole brain lysates of two models. Blots

represent three (sleep-deprived) or two (*Sleepy*) independent experiments. Quantitative analysis of immunoblots of all 14 phosphorylation-motif antibodies is shown in Fig. 1c. n = 12 (S6), 9 (SD6, RS3), 6 (wild-type, *Sleepy*).



Extended Data Fig. 3 | Analysis of sleep phenotype and signalling changes after food-and-water deprivation in the baseline and sleep deprivation conditions. \mathbf{a} - \mathbf{c} , Analysis of circadian (\mathbf{a}) and mean (\mathbf{b}) absolute NREMS delta power, and duration (\mathbf{c}) of NREMS, REMS and wake states, of wild-type mice (n=8) without (sham) or with 6 h of food-and-water deprivation (FD 6 h). \mathbf{d} , Quantitative analysis of immunoblots with six phosphorylation-motif antibodies using whole brain lysates of sham and 6-h food-and-water deprived mice (n=8) collected at ZT6. \mathbf{e} - \mathbf{g} , Analysis of circadian (\mathbf{e}) and mean (\mathbf{f}) absolute NREMS delta power,

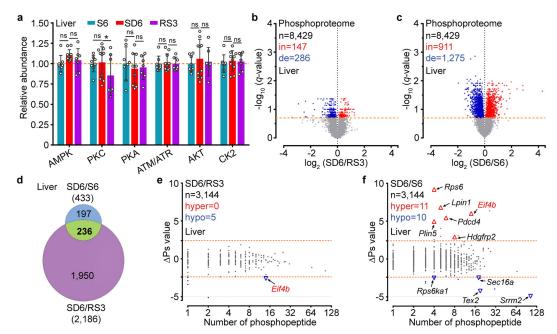
and duration (**g**) of NREMS, REMS and wake states, of wild-type mice (n=11) without (SD + sham) or with 6-h food-and-water deprivation during 6-h sleep deprivation (SD + FD 6h). **h**, Quantitative analysis of immunoblots with six phosphorylation-motif antibodies using whole brain lysates of SD + sham and SD + FD mice (n=6) collected at ZT6. Mean \pm s.e.m., two-way ANOVA, Sidak's test (**a**, **c**, **e**, **g**); Paired t-test, two-tailed (**b**, **f**); Mean \pm s.d., two-way ANOVA, Fisher's LSD test (**d**, **h**). *(black), P < 0.05; ns, P > 0.05.



Extended Data Fig. 4 | See next page for caption.

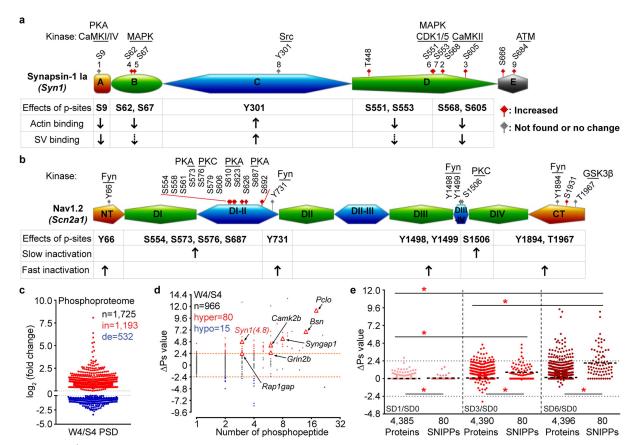
Extended Data Fig. 4 | Quality assessment of proteomic and phosphoproteomic analysis. a, Representative TMT quantification spectrum for the pS551-containing phosphopeptide from the skipped Sik3 exon-13 among phosphoproteomic data of the Sleepy model (two independent experiments). b-e, Quality assessment of one proteomic dataset (EX4, SlpWTpa2) by two search pipelines. Global distribution of protein quantification using Proteome Discoverer (PD v.2.1; n = 8,273) (b) and JUMP (v.1.12.1; n = 8,473) (c). Boxes correspond to the 25th, 50th and 75th percentiles of the data, whiskers extend to 1.5-fold of the interquartile range. A similar number of accepted proteins (1% FDR) were identified by two pipelines (d). Pearson correlation between the two pipelines was calculated for each PSM from quantified proteins by both pipelines (e). The vast majority (99.88%) of PSMs (n = 73,454) have R^2 values larger than 0.9 (red dashed line). f, A Venn diagram showing overlaps of quantified proteins between whole brain proteomes of Sleepy and sleep-deprived models. g, Volcano plots showing comparative

analysis of Sleepy/wild-type, SD6/RS3 and SD6/S6 proteomes. Multiple unpaired t-test (P value) followed by FDR (Q value) analysis. x axis, $\log_2(\text{fold change})$ in abundance; y axis, $-\log(Q$ value) of abundance change. The numbers of total (n), increased (in: Q < 0.2, red) and decreased (de: Q < 0.2, blue) subjects are shown. Orange dotted lines indicate Q = 0.2. \mathbf{h} , Pearson correlation between normalized and unnormalized phosphopeptides in Sleepy/wild-type, SD6/RS3, SD6/S6 groups. The numbers of phosphopeptides that can be normalized are shown. \mathbf{i} , Immunoblots were performed with phosphorylation-site specific antibodies to verify hyper-phosphorylation of several proteins in two models. Three or two independent experiments for sleep-deprived or Sleepy models, respectively. \mathbf{j} , Quantitative analysis of immunoblots in \mathbf{i} , normalized with whole protein abundance, for Sleepy (n = 6) and sleep-deprived (n = 9) models. Mean \pm s.d., two-way ANOVA with Fisher's LSD test. *(black), P < 0.05; *(cyan), P < 0.01; *(red), P < 0.001; ns, P > 0.05.



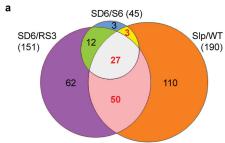
Extended Data Fig. 5 | Liver phosphoproteome analysis of the sleep-deprived model. a, Quantitative analysis of immunoblots with six phosphorylation-motif antibodies using whole liver lysates from the sleep-deprived model. n=8 (S6), 10 (SD6), 7 (RS3). Mean \pm s.d., two-way ANOVA with Fisher's LSD test. *(black), P < 0.05; ns, P > 0.05. b, c, Volcano plots showing comparative analysis of liver phosphoproteomes in the SD6/RS3 (b) and SD6/S6 (c) groups. Multiple unpaired t-test (P value) followed by FDR (Q value) analysis. x axis, $\log_2(\text{fold change})$ in

abundance; y axis, $-\log(Q$ value) of abundance change. The numbers of total (n), increased (in: Q < 0.2, red) and decreased (de: Q < 0.2, blue) subjects are shown. Orange dotted lines indicate Q = 0.2. d, A Venn diagram showing overlaps of significantly changed (Q < 0.2) phosphopeptides among the SD6/RS3 and SD6/S6 groups. e, f, Global Δ Ps analysis of all phosphoproteins identified in the SD6/RS3 (e) and SD6/S6 (f) groups of liver phosphoproteomes. Dotted lines, Δ Ps = ± 2.4 .



Extended Data Fig. 6 | Examples of cumulative phosphorylation of SNIPPs and synaptic phosphoproteomic analysis of normal sleep—wake model. a, b, A schematic of the domain structure of synapsin-1²² (a) and Nav1.2^{23,41,42} (b) that summarizes known phosphorylation sites, kinases and physiological functions. Synapsin-1 can be divided into five domains (domains A–E). Nav1.2 can be divided into cytoplasmic N-terminal (NT), C-terminal (CT), four homologous transmembrane domains (DI–DIV) and intracellular loops (DI–II, DII–III, DIII–IV). Amino acid numbers refer to the sequence of the mouse proteins. Sites 1–9 of synapsin-1 are designated according to the consensus in the literature. Phosphorylation sites that are undetected or unchanged in our experiments are labelled in

grey, whereas those that exhibit significantly increased phosphorylation with sleep deprivation are shown in red. Dashed arrows indicate the presence of contrasting data for biological functions in the literature. c, Published forebrain PSD phosphoproteome results were used for comparative analysis between normal wake (W4) and sleep (S4) brains. d, Global Δ Ps analysis of all identified phosphoproteins in the W4/S4 group. Dotted lines (Δ Ps = \pm 2.4). e, Quantitative Δ Ps analysis of SD1/SD0, SD3/SD0 and SD6/SD0 groups. Mean; one-way ANOVA, Tukey's test (total, SNIPPs); unpaired t-test, two-tailed (total versus SNIPPs). *(red), P < 0.001.



Total: 267 Hyper-Phosphoproteins

v			
	Action	Potential	

Gene	Molecular and Neuronal Functions
Scn1a	Voltage-gated Na+ channel

Voltage-gated Na+ channel; Action potential backpropagation * Scn2a1

Neurotran	eurotransmitter Release		
Gene	Molecular and Neuronal Functions		
Arfgap3	GAPs of ARF1 for endocytosis		
Brsk1	Protein kinase; AMPK-related; Short-term plasticity		
* <u>Bsn</u>	Active zone scaffolding protein; Short-term plasticity		
Cacna1e	Voltage-gated Ca2+ channel		
<u>Cadps</u>	Synaptic vesicle protein for exocytosis; Short-term plasticity		
Camk2b	Protein kinase		
Dmxl2	Scaffolding protein for exocytosis		
* <u>Dnm1</u>	GTPase for endocytosis; Short-term plasticity		
* <u>Pclo</u>	Active zone scaffolding protein		
* <u>Rims1</u>	Active zone protein for exocytosis; Short-term plasticity		
Rims2	Active zone protein for exocytosis		

Synaptic vesicle protein; Short-term plasticity

Dendrite Morphogenesis

* <u>Syn1</u>

	Deliunie ii	noi priogeriesis		
	Gene	Molecular and Neuronal Functions		
	<u>Abi1</u>	Regulator of protein kinase ABL1		
	Arhgap39	GAPs of Rac1 and Cdc42		
	Arhgef2	GEFs of RhoA; AMPAR complex		
*	Mark2	Protein kinase; AMPK-related		
	Mink1	Protein kinase; Rap2 effector		
	Rap1gap	GAPs of Rap1		
*	Sipa1I1	GAPs of Rap2; PSD-95/NMDAR complex		
	Tanc2	PSD scaffolding protein; PSD-95 complex		
	<u>Tnik</u>	Protein kinase; Rap2 effector		

Neurogen	Neurogenesis		
Gene	Molecular and Neuronal Functions		
Camsap1	Microtubule organization; Spectrin-binding		
Clasp2	Microtubule dynamics; Neuronal polarity; +TIPs		
Dock7	GEFs of Rac1 and Rac3; Neuronal polarity		
<u>Gprin1</u>	Gαo binding protein; Cdc42 complex		
* <u>Mark1</u>	AMPK-related kinase; Neuronal polarity		
<u>Nav1</u>	Microtubule dynamics; Neuronal migration; +TIPs		
Rap1gap2	GAPs of Rap1; Axonogenesis		
* Rapgef2	GEFs of Rap and Ras; Axonogenesis		
Trio	Microtubule dynamics: GEFs of Rac1 and RhoG: +TIP		

Extended Data Fig. 7 | Physiological functions of 80 SNIPPs. a, A Venn diagram showing overlaps of the set of hyperphosphorylated proteins ($\Delta Ps > 2.4$) between sleep-deprived and *Sleepy* models. **b**, A summary

	Synaptic Plasticity	
	Gene	Molecular and Neuronal Functions
	Abl2	Protein kinase
	Agap2	GAPs of ARF1 and ARF5
	Ank3	Scaffolding protein
	<u>Anks1b</u>	PSD scaffolding protein
	Baiap2	Adapter protein of Cdc42; Actin reorganization
	Cdkl5	Protein kinase
	Cnksr2	Regulator of protein kinase RAS
	Dlgap2	PSD scaffolding protein
	<u>Dlgap3</u>	PSD scaffolding protein
	<u>Grin2b</u>	Glutamate receptor ionotropic; NMDAR subunit
	<u>Grm5</u>	Glutamate receptor metabotropic
	<u>lqsec1</u>	GEFs of ARF1 and ARF6
	Iqsec 2	GEFs of ARF
	<u>Lrrc7</u>	PSD scaffolding protein
	<u>Mff</u>	Mitochondrial fission
	Plppr4	Lipid phosphatase
	Rab11fip5	Rab effector; Protein trafficking
	Shank2	PSD scaffolding protein
*	: <u>Shank3</u>	PSD scaffolding protein
	Sorbs2	Protein kinase ABL regulator
*	Srcin1	Protein kinase SRC regulator
	Syngap1	GAPs of Ras and Rap; NMDAR complex
	<u>Synpo</u>	Cytoskeleton organization; Actin-binding
	Undetermined	
	Gene	Molecular and Neuronal Functions
	2010300C02Rik	Unknown
	Ankrd63	Unknown
	<u>Arfgap2</u>	GAPs of ARF1; Protein transport
	<u>Arhgap21</u>	GAPs of RhoA and Cdc42; Golgi structure
	C2cd4c	Phospholipid binding
	Caskin1	Active zone scaffolding protein: CASK complex

Caskin1 Active zone scaffolding protein; CASK complex Cep170 Microtubule organization; Centrosomal protein Cep170b Microtubule organization; Centrosomal protein

Clasp1 Microtubule dynamics;

Ddx3x RNA helicase

Ddx3y RNA helicase

Elfn2 Regulator of protein phosphatase PP1

Protein exocytosis Exoc1

<u>Ildr2</u> ER stress; Lipid homeostasis

<u>Мар2</u> Microtubule stiffening

Protein kinase; AMPK-related; Microtubule-associated Mark4

Protein kinase; Microtubule-associated Mast1

Osbpl6 Lipid transport

Pde4b cAMP phosphodiesterase

Pdha1 Mitochondria pyruvate dehydrogenase

Lipid PtdIns transfer Pitpnm2

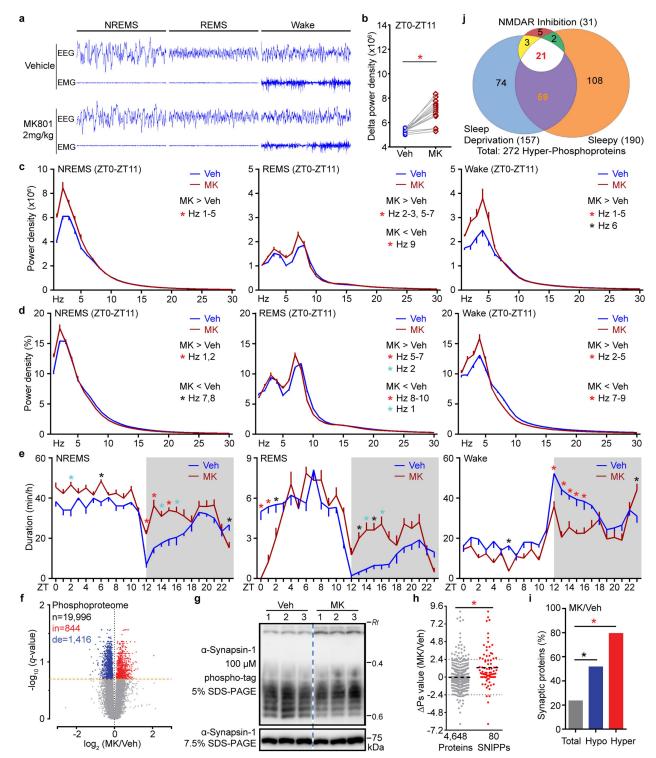
Sphkap Regulator of protein kinase PKA

* Stk32c Protein kinase

GAPs of Rab3A, Rab22A, Rab27A, and Rab35 Tbc1d10b

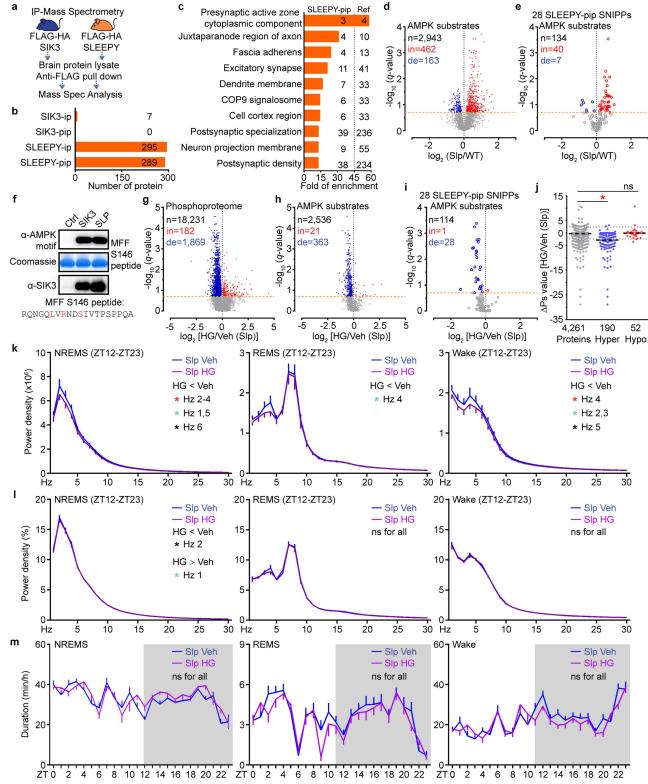
Protein deubiquitination <u>Usp31</u>

of 80 SNIPPs and their physiological functions. Stars mark the 13 SWA-SNIPPs (Fig. 3f). Gene names for annotated synaptic proteins are shown in bold.



Extended Data Fig. 8 | Phosphorylation-state changes of SNIPPs correspond to changes of sleep need in NMDAR inhibition model. a, Representative 8-s EEG and EMG from ZT0–ZT3 for NREMS, REMS and wake for vehicle or MK801-treated mice. b, Mean absolute NREMS delta power analysis of vehicle or MK801-injected mice (n=14). Paired t-test, two-tailed. c-e, Analysis of absolute EEG power spectra (c), relative EEG power spectra (d) and duration (e) for vehicle or MK801-injected wild-type mice (n=14). Mean \pm s.e.m., two-way ANOVA with Sidak's test. f, Volcano plot showing comparison between phosphoproteomes of MK801 and vehicle treated mice. Orange dotted line, Q = 0.2. Multiple unpaired t-test (P value) followed by FDR (Q value) analysis.

g, Phosphorylation state of synapsin-1 was assessed by SDS–PAGE followed by phospho-tag (top) and immunoblotting with anti-synapsin-1 antibody (bottom). The Rf value of 1.0 is defined as the position of bromphenol blue dye (two independent experiments). **h**, Quantitative ΔPs analysis of MK801/vehicle group. Mean, unpaired t-test, two-tailed. **i**, Percentage of synaptic proteins among the total, hypophosphorylated and hyperphosphorylated proteins in the MK801/vehicle group. χ^2 test, two-sided. **j**, Venn diagram showing overlaps of hyperphosphorylated proteins ($\Delta Ps > 2.4$) among all three (*Sleepy*, SD and MK801) models. *(black), P < 0.05; *(cyan), P < 0.01; *(red), P < 0.001.



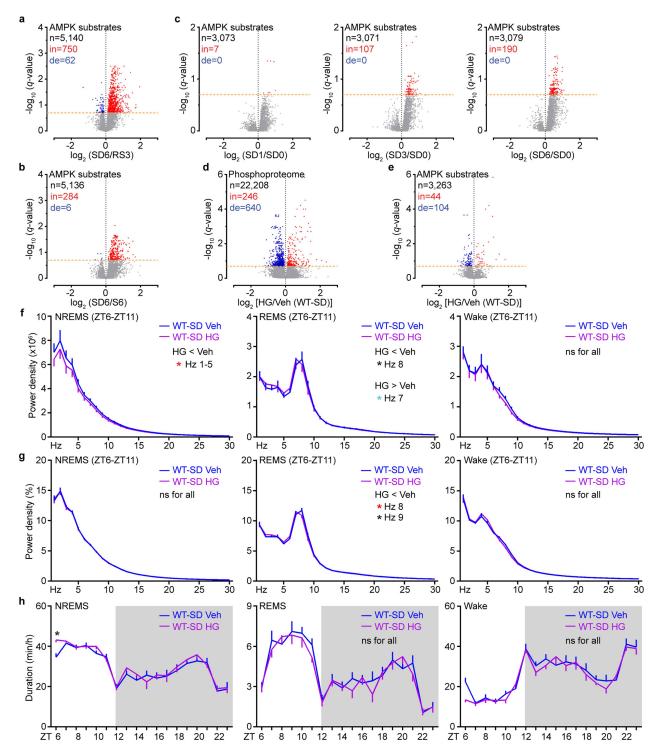
Extended Data Fig. 9 | See next page for caption.



Extended Data Fig. 9 | SLEEPY causes constitutively high sleep need by preferentially associating with and phosphorylating SNIPPs.

a, Experimental design for comparing the interactomes of SIK3 and SLEEPY from whole brain lysates. b, Summary of SIK3 and SLEEPY interacting proteins (ip) and preferential interacting proteins (pip). c, Gene-annotation enrichment analysis of 289 SLEEPY-preferential interacting proteins (SLEEPY-pip). GO cellular component enrichment analysis using all 22,262 genes of *Mus musculus* as reference (Ref). Fisher's exact with FDR multiple test correction was used to determine statistical significance. Top 10 GO terms of fold enrichment (FDR <0.0001), the gene number of SLEEPY-pip and Ref in each term are shown. d, e, Volcano plots showing phosphorylation changes of all putative AMPK substrates in the *Sleepy*/wild-type group (d) or from the 28 SLEEPY-pip SNIPPs (e). Orange dotted lines, Q = 0.2. f, In vitro kinase assay of recombinant

SLEEPY and SIK3, and immunoblotting with AMPK phosphorylation motif antibody (two independent experiments). **g-i**, Volcano plot showing comparative analysis of whole brain phosphoproteomes (**g**), all putative AMPK substrates (**h**) or from 28 SLEEPY-pip SNIPPs (**i**) in the HG/vehicle (Slp) group. Orange dotted lines, Q = 0.2. **j**, Quantitative Δ Ps analysis of 190 hyperphosphorylated proteins and 52 hypophosphorylated proteins in the HG/vehicle (Slp) group. Dotted lines, Δ Ps = ± 2.4 . **k-m**, Analysis of absolute EEG power spectra (**k**), relative EEG power spectra (**l**) and duration (**m**) of NREMS, REMS and wake states of $Sik3^{Slpl+}$ (Slp, n = 14) mice injected with vehicle (Veh) or 8 mg/kg HG at ZT6 and ZT9. Multiple unpaired t-test (P value) followed by FDR (Q value) analysis (**d**, **e**, **g-i**). Mean, one-way ANOVA with Dunnett's test (**j**). Mean \pm s.e.m., two-way ANOVA with Sidak's test (**k-m**). *(black), P<0.05; *(cyan), P<0.01; *(red), P<0.001; ns, P>0.05.



Extended Data Fig. 10 | Inhibition of SIK3 kinase activity reduces phosphorylation of AMPK substrates in sleep-deprived wild-type brains. a-c, Volcano plots showing phosphorylation changes of all putative AMPK substrates in the SD6/RS3 (a), SD6/S6 (b) and time-course sleep-deprivation groups (c). Orange dotted lines, Q = 0.2. d, e, Volcano plots showing comparative analysis of whole brain phosphoryteome (d) and phosphorylation changes of all putative AMPK substrates (e) in the

HG/vehicle (WT-SD) group. Orange dotted lines, Q=0.2. **f-h**, Analysis of absolute EEG power spectra (**f**), relative EEG power spectra (**g**) and duration (**h**) of NREMS, REMS and wake states of sleep-deprived (ZT0–ZT6) wild-type (n=16) mice injected with vehicle (Veh) or 8 mg/kg HG at ZT0 and ZT3. Multiple unpaired t-test (P value) followed by FDR (P value) analysis (P value). Mean P s.e.m., two-way ANOVA with Sidak's test (P value), P value, P value, P value, P value) of the Sidak's test (P value), P value, P valu



Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis

Alexander Crits-Christoph^{1,2}, Spencer Diamond³, Cristina N. Butterfield³, Brian C. Thomas³ & Jillian F. Banfield^{2,3,4,5}*

In soil ecosystems, microorganisms produce diverse secondary metabolites such as antibiotics, antifungals and siderophores that mediate communication, competition and interactions with other organisms and the environment^{1,2}. Most known antibiotics are derived from a few culturable microbial taxa³, and the biosynthetic potential of the vast majority of bacteria in soil has rarely been investigated⁴. Here we reconstruct hundreds of near-complete genomes from grassland soil metagenomes and identify microorganisms from previously understudied phyla that encode diverse polyketide and nonribosomal peptide biosynthetic gene clusters that are divergent from well-studied clusters. These biosynthetic loci are encoded by newly identified members of the Acidobacteria, Verrucomicobia and Gemmatimonadetes, and the candidate phylum Rokubacteria. Bacteria from these groups are highly abundant in soils⁵⁻⁷, but have not previously been genomically linked to secondary metabolite production with confidence. In particular, large numbers of biosynthetic genes were characterized in newly identified members of the Acidobacteria, which is the most abundant bacterial phylum across soil biomes⁵. We identify two acidobacterial genomes from divergent lineages, each of which encodes an unusually large repertoire of biosynthetic genes with up to fifteen large polyketide and nonribosomal peptide biosynthetic loci per genome. To track gene expression of genes encoding polyketide synthases and nonribosomal peptide synthetases in the soil ecosystem that we studied, we sampled 120 time points in a microcosm manipulation experiment and, using metatranscriptomics, found that gene clusters were differentially co-expressed in response to environmental perturbations. Transcriptional co-expression networks for specific organisms associated biosynthetic genes with two-component systems, transcriptional activation, putative antimicrobial resistance and iron regulation, linking metabolite biosynthesis to processes of environmental sensing and ecological competition. We conclude that the biosynthetic potential of abundant and phylogenetically diverse soil microorganisms has previously been underestimated. These organisms may represent a source of natural products that can address needs for new antibiotics and other pharmaceutical

We reconstructed draft genomes for hundreds of microorganisms from the soil ecosystem of a northern Californian grassland using genome-resolved metagenomic methods, and targeted genomes from four dominant soil phyla for analysis of their biosynthetic potential (Extended Data Fig. 1). Specifically, we analysed newly reconstructed genomes from 149 Acidobacteria, 135 Verrucomicrobia, 43 Rokubacteria and 49 Gemmatimonadetes species (Supplementary Table 1 and Supplementary Methods). We targeted these groups because bacteria from all four phyla are highly abundant at our field sampling site⁸ (Fig. 1a) and in globally sampled soils⁵. Specifically, meta-analysis of many 16S rRNA gene sequence studies showed that Acidobacteria and Verrucomicrobia are the first and second most abundant bacterial phyla in soil, respectively⁵, and Gemmatimonadetes

are also known to be common in soils⁹. There are few reference genomes available for soil-associated bacteria from all four phyla, and their potential for secondary metabolism remains understudied. To our knowledge, the current study represents the largest genomic sampling of soil-associated bacteria from these groups to date and the most detailed analysis of their secondary metabolism.

Within the genomes, we identified 1,159 biosynthetic gene clusters on contigs at least 10 kb in length (Fig. 1b and Supplementary Table 2) and an additional 440 biosynthetic gene clusters on smaller contigs (Supplementary Table 3) using antiSMASH 3.0¹⁰, an in silico pipeline that was originally verified against 473 verified biosynthetic gene clusters with a 97.7% reported accuracy¹¹. The gene clusters that we identified are inferred to synthesize nonribosomal peptides (NRPs), polyketides, terpenes, bacteriocins, lassopeptides, lantipeptides and metabolites of uncertain function. Most known bacterial natural products—including many of the clinical antibiotics that we use today have been obtained from microbial isolates³ of the Actinobacteria, Proteobacteria and Bacillus, which represent microorganisms that often comprise a minority in soil microbial communities^{4,5}. Previous global analyses based on the few publicly available genomes for Acidobacteria, Verrucomicrobia and Gemmatimonadetes¹²⁻¹⁴ identified only a handful of biosynthetic clusters, and to our knowledge only the Acidobacteria have previously been suggested to be linked to secondary metabolite production^{7,15}. We greatly expand the number of known biosynthetic gene pathways from these soil microorganisms and at the same time confidently link them to their genomic contexts.

Most previous searches for biosynthetic systems from uncultivated microorganisms have randomly cloned environmental DNA into a host organism to screen for function (functional metagenomics)¹⁶. Other studies^{2,17} have used degenerate PCR primers to explore the genetic diversity of novel biosynthetic clusters without the need for cloning, but primers can fail to amplify genetically divergent sequences. Because we reconstructed near-complete genomes de novo, we could identify entire novel biosynthetic gene clusters as well as describe their genomic, phylogenetic and ecological contexts within individual genomes and the environment. We computationally tested the ability of sets of previously used degenerate primers^{2,17} to detect genes containing polyketide ketoacyl synthase and NRP amino acid adenylation domains in the clusters reported here, and found that only 5 out of 240 clusters would be likely to amplify properly when using degenerate primers (Supplementary Table 6).

Gene clusters containing nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs) were of particular interest, as the products of these enzymes include many antibiotics, antifungals, siderophores and immunosuppressants¹⁴. These NRPS and PKS biosynthetic pathways use modular enzymatic domains to build molecules with complex chemical structures. We identified 240 NRPS, PKS (types I, II and III, which differ in the organization of their enzymatic domains) and hybrid (NRPS-PKS) gene clusters on contigs from all four phyla of interest (Fig. 1c and Supplementary Table 4) and 86 probably incomplete clusters on smaller genome fragments. Although they

¹Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA, USA. ²The Innovative Genomics Institute, University of California, Berkeley, CA, USA. ³Department of Earth and Planetary Science, University of California, Berkeley, CA, USA. ⁴Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, USA. *e-mail: jbanfield@berkeley.edu

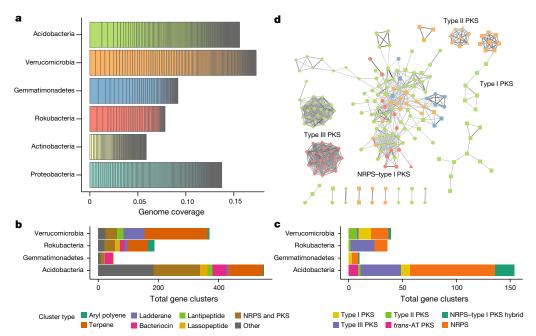


Fig. 1 | Diversity of extracted soil genomes and their biosynthetic gene clusters. a, Mean relative abundances of reconstructed genomes across 60 soil samples as determined by sequencing coverage of the genomes. Genomes from four understudied soil phyla are juxtaposed with recovered genomes from the Actinobacteria and Proteobacteria for comparison. b, Biosynthetic gene clusters found on contigs greater than 10 kb, from

each phylum studied, coloured by putative product types as assigned by antiSMASH. **c**, NRPS and PKS gene clusters found on contigs >10 kb, from each phylum studied. **d**, Network of biosynthetic gene clusters, in which edges connect clusters that share genes. The line thickness and darkness increase with increasing percentage of genes shared between clusters. *trans*-AT, *trans*-acyltransferase.

are enormously diverse in gene content, these biosynthetic pathways are identifiable owing to their colocalized logical organization of conserved enzymatic domains. Although the majority of these clusters occurred in a wide diversity of Acidobacteria, we also identified 11 NRPS clusters in genomes of the Rokubacteria, a recently described phylum that was not previously known to produce natural products. The co-linear 'assembly-line' regulation of many NRPS and type I PKS systems make predictions of the core scaffold of the molecular product synthesized possible ^{11,18}. In 136 cases, there were a sufficient number of functional domains with known substrate specificity to predict the core chemical structures of the products using antiSMASH (Supplementary Table 4).

To compare the degrees to which predicted biosynthetic clusters shared genes, we built a relational network of clusters on the basis of shared gene content. This approach revealed substantial genetic variety, with large groups of diverse and sparsely connected NRPS and PKS systems in Verrucomicrobia, Acidobacteria and Rokubacteria and many unique NRPS-based clusters with few close representatives (Fig. 1d). A conserved type III PKS locus that was nearly ubiquitous in the Rokubacteria formed a dense network cluster, as did a conserved type III PKS locus found in a wide clade of the Acidobacteria. The high conservation of these type III PKS loci across taxonomic groups could indicate a broad distribution of a novel group of specialized metabolites.

We compared the 240 NRPS and PKS gene clusters to the reference set described in the 'Minimum Information about a Biosynthetic Gene' (MIBiG) repository (Supplementary Table 5). No protein in any cluster shared with reference proteins more than 79.7% amino acid identity across \geq 50% of the full protein lengths. Fifty-nine per cent of predicted proteins had no \geq 50%-length homologue in MIBiG, and those that did shared an average of only about 39% amino acid identity to the best hit of any MIBiG protein. Using the same thresholds for gene homologues, we found that 220 clusters did not share more than 50% of the genes of any previously described cluster. Although the relationship between gene similarity of biosynthetic genes and structural similarities of their final products can be difficult to discern, previous analyses have shown that structural divergence correlates strongly with genetic divergence, even within families of gene clusters 20.

It is often the case that antibiotic producers will also encode antibiotic resistance genes to avoid self-toxicity, and that these genes will often co-localize with the antibiotic biosynthetic cluster in the genome²¹. Therefore, the presence of antimicrobial resistance genes within a gene cluster could indicate that the cluster is involved in antibiotic production. We mined all NRPS and PKS biosynthetic loci with a set²² of curated hidden Markov models for antibiotic resistance proteins (in part derived from the Resfams²³ database) (Supplementary Methods). One hundred and fifty-three proteins from 84 different NRPS and PKS clusters most closely matched hidden Markov models for transporters known to be involved in antimicrobial resistance, out of a total of 621 transporter genes within clusters. Annotations that could most confidently be linked to antibiotic resistance included one D-alanine-D-alanine ligase in a Rokubacteria NRPS cluster, four D-alanine-D-alanine ligases in acidobacterial NRPS clusters, and two modified penicillin-binding protein sequences in Verrucomicrobia NRPS clusters (Supplementary Table 7).

Two near-complete genomes of divergent Acidobacteria were found to encode unusually large repertoires of NRP and PKS gene clusters. We refer to these two organisms as 'Candidatus Eelbacter' (genome Eelbacter_gp4_AA13) and 'Candidatus Angelobacter' (genome Angelobacter_gp1_AA117), tentatively placed within the Blastocatellia and the Acidobacteriales, respectively. In the 7-Mb genome of Candidatus Eelbacter we identified 17 biosynthetic loci containing 74 NRPS and PKS open reading frames that were 404 kb in total length. In the 6.5-Mb genome of Candidatus Angelobacter there were 16 loci containing 54 NRP/PKS open reading frames that were 325 kb in total length. The biosynthetic genes from each species had only distant homology to those from the other. We confirmed the biosynthetic clusters for both genomes by re-analysing with 'Prediction Informatics for Secondary Metabolomes' (PRISM)²⁴ (Extended Data Figs. 2, 3). In total, each of these organisms contains over 900 kb of genes that are putatively involved in biosynthesis of secondary metabolites (about 12–14% of their recovered genomes). A phylogenetic analysis, using ribosomal protein sequences, of acidobacterial genomes from this study and reference databases revealed that both Candidatus Angelobacter

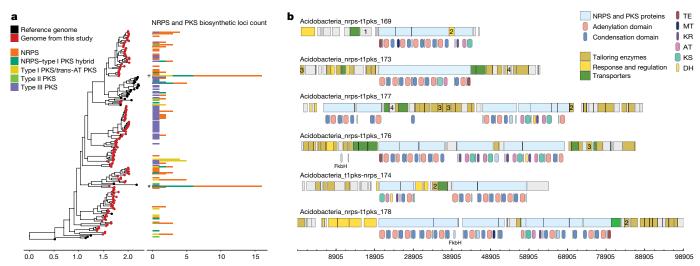


Fig. 2 | Biosynthetic NRPS and PKS loci from the Acidobacteria.
a, Concatenated ribosomal protein phylogenetic tree of all acidobacterial genomes from this study (red) and existing reference genomes (black). Scale bar on the tree represents substitutions per site. Adjacent is a chart that reflects the count of NRPS and PKS biosynthetic gene clusters observed in each genome. The phylogenetic placements of *Candidatus* Eelbacter (*) and *Candidatus* Angelobacter (+) are marked. b, Six large PKS-NRPS hybrid biosynthesis gene clusters are encoded in the

Candidatus Eelbacter genome. Predicted genes and biosynthetic protein domains are coloured by general function, and the genomic positions of polyketide and nonribosomal peptide synthetic domains are shown below each genome track. The following gene annotations are identified by number: 1, penicillin amidase; 2, oxygenase; 3, radical SAM proteins; and 4, betalactamase. AT, acyltransferase; DH, dehydrogenase; KR, ketoreductase; KS, ketosynthase; MT, methyltransferase; TE, thioesterase.

and *Candidatus* Eelbacter acquired their unusual arrays of biosynthetic operons independently in evolutionary time (Fig. 2a).

The Candidatus Angelobacter genomes included multiple lantibiotic biosynthesis proteins, a bacteriocin biosynthesis cluster, multigene operons with components for both a type VI and a type II secretion system, and several large RHS-repeat containing proteins, which have been hypothesized to have evolved to mediate microbial competition by facilitating transfer of protein toxins between species²⁵. The Candidatus Eelbacter genome contained six clusters that were complex type I NRPS-PKS hybrid systems over 45 kb in length (Fig. 2b). Three replicate genomes of Candidatus Eelbacter were obtained from independent soil samples and shared the same set of biosynthetic clusters. Both species also possessed CRISPR-Cas loci (31 spacers and repeats in Candidatus Angelobacter and 438 across the *Candidatus* Eelbacter genome). The ecological and evolutionary forces that can select for the production of an unusually high number of metabolites in a species are varied, and previously characterized examples are microorganisms with complex cooperative lifestyles^{26,27} or an association with a eukaryotic host²⁸. The discovery of these two microorganisms establishes that bacterial specialization in secondary metabolite biosynthesis is not limited to known clades in the Actinomycetales, Proteobacteria, Cyanobacteria, Bacilli and the recently discovered Entotheonella²⁸. When considered together, the genomic features of these Acidobacteria hint towards an unusually competitive lifestyle mediated by chemical and toxin production.

We tested whether the microorganisms genomically described in this study are active and express biosynthetic NRPS or PKS gene clusters by analysing metatranscriptomics data from 120 soil microcosm samples from two soil depths and two sampling locations from the same field site that were subject to amendment with glucose, methanol or water over 24 h (Supplementary Methods). These experiments were designed to probe the strong biological responses that occur in soils following water addition and nutrient release after a long dry period²⁹. Because distinct NRPS or PKS clusters can produce products with very different bioactivities, we tracked expression of each gene cluster as a functional biosynthetic unit by pseudo-aligning exact matches of paired reads to full genomes obtained directly from the environment studied using Kallisto³⁰. Overall, we detected expression for 198 NRPS and/or PKS genes across those NRPS and PKS clusters with any level of gene expression (133 out of 180 clusters) (Supplementary Table 8). Expression of NRPS and PKS clusters was detected in all four phyla that we studied,

and 84 active clusters were detected in Acidobacteria (Extended Data Fig. 4). We detected the expression of genes within 10 biosynthetic clusters—including 11 genes with NRPS and/or PKS domains within these clusters—of *Candidatus* Eelbacter (Extended Data Fig. 5) and 14 clusters of *Candidatus* Angelobacter—including 25 genes with NRPS and/or PKS domains. We tested for co-expression of genes in all biosynthetic clusters and found that gene clusters were co-expressed more often than were randomized permutations of genes across each genome (Wilcoxon rank-sum test, P < 0.001).

Across all organisms in our dataset, we identified ten NRPS and/ or PKS gene clusters from seven genomes with levels of expression that were time-dependent across the 24-h time course of the amendment experiments (permutational multivariate analysis of variance (PERMANOVA); P < 0.05, false discovery rate (FDR) = 5%) (Fig. 3a) and Extended Data Fig. 6). We confirmed differential expression over time for individual genes within these clusters using a model that accounts for variation in both sequencing library sizes and organism abundances across samples³¹ (DESeq 2^{32} ; P < 0.05; FDR = 5%) (Supplementary Table 9). Notably, the expression of genes from several gene clusters in Candidatus Angelobacter showed a statistically significant increase 12-24 h after substrate addition (Fig. 3a), and we found that the expression of several biosynthetic genes of Candidatus Angelobacter was temporally distinct from the expression of core ribosomal genes (Fig. 3b). These results indicate that Candidatus Angelobacter populations respond to water and substrate addition, and independently regulate expression of secondary metabolite genes many hours after a period of increased core metabolic gene expression.

To predict the broader biological and ecological roles of these biosynthetic NRPS and PKS genes, we conducted separate co-expression analyses of all genes for each of the seven species identified with temporally dependent biosynthetic gene expression, using the WGCNA package³³ (Supplementary Methods), across the 120 microcosm timepoint samples. Co-expressed genes often share biological functions and regulation³⁴. Modules of co-expressed genes significantly enriched in secondary metabolite genes were identified in four out of seven genomes (P < 0.05; hypergeometric distribution) (Fig. 3c, Extended Data Fig. 7 and Supplementary Table 10). These four modules were small (fewer than 69 genes) and very transcriptionally distinct. We found that all four secondary metabolism networks were dominated by genes involved in two-component systems, efflux and transcriptional

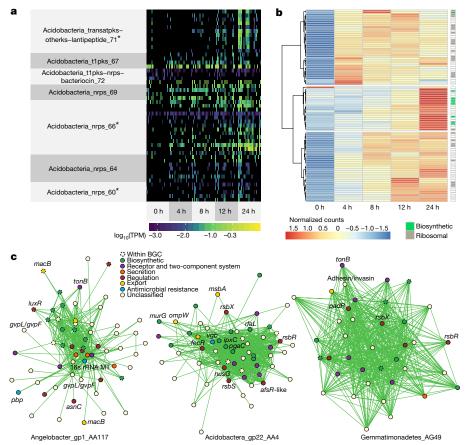


Fig. 3 | **Metatranscriptomics of biosynthetic genes. a**, Levels of transcriptional expression of genes from biosynthetic gene clusters encoded in the *Candidatus* Angelobacter genome, across 120 microcosm soil samples grouped by extraction times (reported in hours). Expression levels are reported in \log_{10} -transformed transcripts per million (TPM). Gene clusters that were significantly differentially expressed across time points (PERMANOVA); $^*P < 0.05$, FDR = 5% are marked by an asterisk. **b**, Hierarchical clustering of expression levels for differentially expressed (n = 120; DESeq2; P < 0.05; FDR = 5%) genes from the *Candidatus* Angelobacter genome across samples grouped by experimental time point. Differentially expressed genes from biosynthetic clusters and differentially

expressed core ribosomal proteins are marked. Values are reported in counts transformed using the rlog transformation from DESeq2 and were normalized by row. **c**, The transcriptional co-expression network modules (n=120 microcosm time-point samples) significantly enriched in NRPS and PKS biosynthetic genes from three genomes (P<0.05; hypergeometric distribution). Nodes represent gene transcripts and edges between them represent high topological overlap values between the transcripts. Genes outlined are genes found within biosynthetic gene clusters (BGC), and are coloured by assigned function using the Kyoto Encyclopedia of Genes and Genomes and Pfam databases. 16s rRNA MT, gene encoding for a 16S rRNA methyltransferase.

regulators, and were almost completely devoid of genes for the core processes of transcription, translation and energy metabolism.

For Candidatus Angelobacter, genes from five biosynthetic clusters were co-expressed together in a module with a variety of genes involved in environmental sensing and response, including homologues of the gene that encodes for the iron siderophore uptake receptor TonB. Homologues of the gene that encodes for the macrolide export transporter MacB were also found to be co-expressed with the biosynthetic genes, as were two putative antimicrobial resistance genes—those encoding for penicillin-binding protein and for a 16S rRNA methyltransferase. Additional co-expressed genes included an operon for a type VI secretion system and an operon annotated as encoding for gas vesicle proteins. Notably, the Angelobacter population expressed biosynthetic genes from multiple clusters simultaneously, suggesting a concerted response that is linked to ecological competition.

Acidobactera_gp22_AA4 was found to co-express its NRPS gene cluster (Acidobacteria_nrps_112) with response-regulatory genes and a set of genes involved in cell surface structure remodelling, as well as an operon of genes involved in regulating stress response (*rsbX*, *rsbR* and *rsbS*). A homologue of virginiamycin B lyase (*vgb*), which is an inactivator of type B streptogramin antibiotics, was also co-expressed in this module. The same operon of genes involved in the regulation of stress response was found to be co-expressed in the transcriptional network containing a biosynthetic cluster (cluster

Gemmatimonadetes_nrps_183) in Gemmatimonadetes_AG49, along with a *tonB* homologue.

In summary, we uncovered extensive evidence for secondary metabolite synthesis in a large collection of bacterial genomes from four phyla of soil bacteria that have not previously been genomically linked to this capacity. Although we cannot confidently predict more than the basic chemical scaffolds of the products derived from the biosynthetic genes reported here, or their biological activities, a large percentage of known polyketide and nonribosomal metabolites isolated from microbial sources have antimicrobial activity³⁵. Transcriptional associations between specific NRPS and PKS gene clusters, regulators of iron metabolism and putative antimicrobial resistance mechanisms suggest that these gene clusters may be involved in competition for iron resources and antibiotic production. The findings underline the utility of genome-resolved metagenomic investigations of soil ecosystems and open the way for laboratory characterization of genes for novel bioactive metabolites with potential ecological and pharmaceutical importance.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0207-y.

Received: 2 August 2017; Accepted: 2 May 2018; Published online 13 June 2018.



- Hibbing, M. E., Fuqua, C., Parsek, M. R. & Brook Peterson, S. Bacterial 1. competition: surviving and thriving in the microbial jungle. Nat. Rev. Microbiol. 8, 15-25 (2010).
- Charlop-Powers, Z., Owen, J. G., Reddy, B. V., Ternei, M. A. & Brady, S. F. Chemical-biogeographic survey of secondary metabolism in soil. Proc. Natl Acad. Sci. USA **111**, 3757–3762 (2014).
- 3. Cragg, G. M. & Newman, D. J. Natural products: a continuing source of novel drug leads. *Biochim. Biophys. Acta* **1830**, 3670–3695 (2013). Rappé, M. S. & Giovannoni, S. J. The uncultured microbial majority. *Annu. Rev.*
- Microbiol. 57, 369-394 (2003).
- Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. Nat. Rev. Microbiol. 15, 579-590 (2017).
- Bergmann, G. T. et al. The under-recognized dominance of Verrucomicrobia in soil bacterial communities. Soil Biol. Biochem. 43, 1450-1455 (2011).
- Kielak, A. M., Barreto, C. C., Kowalchuk, G. A., van Veen, J. A. & Kuramae, E. E. The ecology of Acidobacteria: moving beyond genes and genomes. Front. Microbiol. 7, 744 (2016).
- Butterfield, C. N. et al. Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. PeerJ 4, e2687 (2016).
- DeBruyn, J. M., Nixon, L. T., Fawaz, M. N., Johnson, A. M. & Radosevich, M. Global biogeography and quantitative seasonal dynamics of Gemmatimonadetes in soil. *Appl. Environ. Microbiol.* **77**, 6295–6300 (2011).
- Weber, T. et al. antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* 43, W237–W243 (2015).
- 11. Medema, M. H. et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 39, W339–W346 (2011).
- 12. Hadjithomas, M. et al. IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. MBio 6, e00932-e15 (2015).
- 13. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 158, 412-421 (2014).
- 14. Wang, H., Fewer, D. P., Holm, L., Rouhiainen, L. & Sivonen, K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. Proc. Natl Acad. Sci. USA 111, 9259-9264
- 15. Parsley, L. C. et al. Polyketide synthase pathways identified from a metagenomic library are derived from soil Acidobacteria. FEMS Microbiol. Ecol. 78, 176-187 (2011).
- 16. Rondon, M. R. et al. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. 66, 2541-2547 (2000).
- 17. Charlop-Powers, Z. et al. Global biogeographic sampling of bacterial secondary metabolism. eLife 4, e05048 (2015).
- 18. Fischbach, M. A. & Walsh, C. T. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. Chem. Rev. 106, 3468-3496 (2006).
- Medema, M. H. et al. Minimum information about a biosynthetic gene cluster. Nat. Chem. Biol. **11**, 625–631 (2015).
- 20. Medema, M. H., et al. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. PLoS Comput. Biol. 10, e1004016 (2014).
- Thaker, M. N. et al. Identifying producers of antibacterial compounds by screening for antibiotic resistance. Nat. Biotechnol. 31, 922–927 (2013).
- Johnston, C. W. et al. Assembly and clustering of natural antibiotics guides target identification. Nat. Chem. Biol. 12, 233-239 (2016).

- 23. Gibson, M. K., Forsberg, K. J. & Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. 9, 207-216 (2015).
- Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Res. 45, W49-W54 (2017).
- Koskiniemi, S. et al. Rhs proteins from diverse bacteria mediate intercellular competition. Proc. Natl Acad. Sci. USA 110, 7032-7037 (2013).
- Claessen, D., de Jong, W., Dijkhuizen, L. & Wösten, H. A. Regulation of Streptomyces development: reach for the sky. Trends Microbiol. 14, 313-319
- Žhang, Y., Ducret, A., Shaevitz, J. & Mignot, T. From individual cell motility to collective behaviors: insights from a prokaryote, Myxococcus xanthus. FEMS Microbiol. Rev. 36, 149-164 (2012).
- Wilson, M. C. et al. An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature 506, 58-62 (2014).
- Unger, S. et al. The influence of precipitation pulses on soil respirationassessing the "Birch effect" by stable carbon isotopes. Soil Biol. Biochem. 42, 1800-1810 (2010).
- Bray, N., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525-527 (2016).
- Klingenberg, H. & Meinicke, P. How to normalize metatranscriptomic count data
- Find the state of the state of
- Langfelder, P & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
- Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. Science 302, 249-255 (2003)
- 35. Bérdy, J. Bioactive microbial metabolites. J. Antibiot. (Tokyo) 58, 1-26 (2005).

Acknowledgements We thank S. Spaulding for assistance with fieldwork, and M. Traxler and W. Zhang for helpful discussions. Sequencing was carried out under a Community Sequencing Project at the Joint Genome Institute. Funding was provided by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy Grant DOE-SC10010566, the Paul G. Allen Family Foundation and the Innovative Genomics Institute of the University of California, Berkeley.

Author contributions A.C.-C. performed genomic and transcriptomic analysis; S.D. performed metagenome assembly and curation; C.N.B. performed microcosm experiments and RNA extractions; A.C.-C., S.D. and J.F.B. wrote the manuscript; B.C.T. supported the metagenomics bioinformatics work; and J.F.B. supervised the project.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-

Supplementary information is available for this paper at https://doi.org/ 10.1038/s41586-018-0207-v

Reprints and permissions information is available at http://www.nature.com/ reprints.

Correspondence and requests for materials should be addressed to J.F.B. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessments.

Soil sampling and DNA extraction. Soil samples were collected from the Angelo Coast Range Reserve meadow (39° 44′ 21.4″ N 123° 37′ 51.0″ W) on four dates in 2014 that bracketed the first winter rain of the season. Samples were collected from three depths, 10–20 cm, 20–30 cm and 30–40 cm at six independent sampling sites that were first metagenomically characterized as part of a previous study⁸. Sampling was conducted in biological triplicate, with three of the sites being unamended biological control plots and three being amended with extended spring rainfall from a sprinkler system as described in a previous publication⁸. Sampling was accomplished using a soil coring device that was fitted with sterilized polycarbonate sheaths. Sheaths were removed after each collection event. After collection, samples were flash-frozen in a mixture of dry ice and ethanol, and placed on dry ice for transport. A total of 60 soil cores were sampled across all depth and treatment conditions.

For each depth, DNA was extracted using MoBio Laboratories PowerMax Soil DNA Isolation kits from 10 g of soil as previously described Mean DNA concentration in the extracted samples, quantified by using qubit fluorometric assay, was 388 ng/ μ l.

Sequencing, genomic assembly and binning. Metagenomic libraries for all 60 samples were prepared and sequenced at the Joint Genome Institute using an Illumina HiSeq 2500 platform to generate 250-bp paired-end reads. Samples were multiplexed for sequencing. Raw sequence data were processed with BBmap³6 to remove Illumina adaptor and phiX sequences, and reads were quality-score trimmed using Sickle with default parameters³7. Read sets were subsequently analysed for per-base GC content using FastQC³8, and it was determined that GC content increased substantially after 200 bp in some sample read sets. Thus all reads longer than 200 bp were hard-trimmed to 200 bp using BBmap. In total, 6.22×10^9 reads were sequenced across all samples, which yielded 1.24 Tb of total sequence information with an average read count of 1.04×10^8 reads per sample.

The 60 samples were individually assembled de novo on a 24-core Intel Xenon Linux cluster node with 256 Gb of RAM using IDBA-UD 39 with the following initial parameters: –pre_correction,—mink 30,—maxk 200,—step 10. In the 13 cases in which assemblies did not complete owing to memory requirements, minimum k-mer size was increased to 40 bp. The resulting assemblies averaged 1.15 Gb of assembled sequence with an N50 of 1,609 bp. Sequencing coverage of each contig was calculated by mapping raw reads back to assemblies using Bowtie2 40 ; 36.4% of reads mapped back to assembled sequence on average. It should also be noted that contigs > 100 kb in length were acquired from all 60 assemblies, with a maximum contig size across assemblies of 2.7 Mb.

All resulting assemblies were subsequently clustered into genome bins individually using a hybrid binning approach. Initially, reads from all assemblies were separately cross-mapped to all scaffolds >2 kb in size from a single assembly using Bowtie2 to generate a coverage profile for the scaffolds of that assembly across all samples. Scaffold differential coverage profiles were used to inform five separate automated binning software packages: ABAWCA, ABAWACA2⁴¹, MaxBin2⁴², ${\rm CONCOCT^{43}}$ and ${\rm MetaBAT^{44}}$, which were run on all samples individually. The resulting output genome bins for all packages run on a single sample were combined, assessed for completeness using an inventory of 51 universal single-copy genes (SCGs), and dereplicated by selecting the most complete bin of an overlapping set using DAStool⁴⁵. Following automated binning, all genomic bins were manually inspected and curated using our in-house bin visualization and analysis system, ggKbase $^{46}\,(\rm http://ggkbase.berkeley.edu).$ Finally, after manual curation in ggKbase, reads from a given sample were mapped back to the bins derived from that sample to identify and correct assembly and scaffolding errors, as previously described⁴⁷. In total, 10,463 individual genome bins were identified across all samples. Of these bins, 3,334 were then estimated at a completeness of \geq 70% using CheckM⁴⁸. Taxonomic assignment of bins was performed by looking at the closest known hits and phylogenetic placement of ribosomal marker proteins. Bins were then dereplicated by clustering their ribosomal S3 proteins at 99% amino acid identity and choosing the bin in each cluster with the highest completeness and lowest contamination, which resulted in a final set of 377 nonredundant bins in the bacterial phyla of interest.

Genomic analysis of genomes and biosynthetic gene clusters. Curated genomes were individually processed using antiSMASH 3.0^{10} with default parameters. The results are summarized in Supplementary Table 2 for gene clusters on contigs greater than 10 kb, Supplementary Table 1 for gene clusters on contigs smaller than 10 kb and Supplementary Table 4 for all PKS and NRPS clusters on contigs greater than 10 kb. Ribosomal protein phylogenetic trees were built using a concatenated set of 16 ribosomal proteins ⁴⁹ for all Acidobacteria genomes in this dataset, as well as those that could be obtained from GenBank or the Integrated Microbial Genomes platform. An *Escherichia coli* genome was used as an outgroup for the

tree. These protein sequences were aligned with MUSCLE 50 and then a maximum likelihood phylogeny was built using FastTree2 51 with default parameters.

To test whether existing primer-based methods have the ability to amplify these biosynthetic gene sequences, sets of forward and reverse degenerate primers used by previous analyses of biosynthetic genetic diversity^{2,17} for ketosynthase genes and adenylation domain genes were searched for pattern matches against all NRPS and PKS clusters in both reverse and forward reading frames. The inosine nucleotides were substituted with the ambiguous code B, because these nucleotides can base pair with adenine, cytosine and uracil. Only five of our gene clusters had correctly oriented matches to both a forward and reverse primer within 2 kb of each other (Supplementary Table 6).

The network of gene clusters based on shared gene content was built by performing an all-versus-all BLASTP search of predicted biosynthetic protein sequences. Shared proteins were defined as protein alignments with at least 50% of the query sequence covered and amino acid per cent identity >50%. Two clusters (nodes) were connected if either one shared at least 10% of its proteins with the other. The width and colour intensity of the network edges was scaled with the length of the shared protein alignments, normalized to the length in base pairs of the two clusters being compared. Biosynthetic gene clusters were compared to clusters previously reported in the MiBIG repository¹⁹ using BLASTP and the same definition of shared proteins, and the closest hits to MiBIG clusters containing at least five genes were reported. To identify antibiotic resistance genes in clusters, we searched protein products of all biosynthetic gene clusters with a set of hidden Markov models derived from a previous publication²², using HMMER with the gathering threshold cutoffs specified in this previous study. We then manually curated hits and eliminated matches to ambiguous functions (acetyltransferases, general methyltransferases and amidases) and focused on reporting proteins with functions that are unlikely to be involved in generic biosynthetic pathways. The Candidatus Angelobacter and Eelbacter genomes were both subsequently analysed using the PRISM3 webserver²⁴.

Soil microcosm experiments and RNA extraction. At the Angelo Coast Range Reserve meadow, five holes were bored within a 1-m² area to obtain 10-cm-long cores of soil, from depths 10-20 cm and 30-40 cm (permission under APP # 27790). Samples were collected on 21 September 2015. At each depth, five cores were mixed in a large Whirl-Pak bag, then distributed into five capped core liners and stored in individual Whirl-Pak bags at 4°C. The unsieved soils were mixed a second time in the laboratory to obtain six equally proportioned samples, and the weights were measured. To settle the soil, the core liners were struck with a rubber mallet 50 times each, and then stored at 4 °C. The night before wet-up experiments, the cores were placed in a cooler alongside the substrate that was to be added, so that the soil and substrate equilibrated to the same temperature and the soil would be kept in the dark. Immediately before adding the substrate, 10 g soil was collected for DNA extraction and 2 g soil with 4 ml LifeGuard RNA Soil Preservation Solution (MoBio) was collected for RNA purification. Both were immediately frozen in liquid N_2 and stored in a freezer at $-80\,^{\circ}\text{C}$. Samples at different time points were collected for nucleic acid extraction in the same manner. Ten millimolar glucose, methanol or water substrate was added to the open-soil core liners and soil in a cooler by pipette 2.5–4 ml at a time over 1 min, and the lid was closed. Substrates were added in amounts that increased the soil moisture to the level of a sample collected from the meadow after 29 cm of rainfall on 5 November 2015 (the moisture level of the field sample was determined by weight loss on drying). RNA was isolated from 2 g soil with RNA PowerSoil Total RNA Isolation kits, following kit protocols. cDNA libraries were prepared and were sequenced to generate 5.9×10^9 150-bp paired-end reads.

Transcriptomics. To test for the expression of clusters of biosynthetic genes within a soil environment, we analysed metatranscriptomics data from experimental soil microcosms. Soil samples from depths of 20 cm and 40 cm from two sampling locations were subject to amendment with glucose, methanol or water, and RNA was extracted from samples at 0, 4, 8, 12 and 24 h after treatment. From the 120 sequenced samples, we generated 5.9×10^9 150-bp paired-end reads. Transcript abundances for all Prodigal-predicted gene sequences from all genomes reconstructed from the project site were quantified using Kallisto³⁰ exact pseudoalignments of paired reads. Kallisto was run using default parameters. Transcripts that were either found to be expressed in at least 10% of samples or to have at least 100 counts were reported and included in downstream analyses. Differential gene expression analysis was performed using PERMANOVA and DESeq2³² (see 'Statistical analysis').

We mapped RNA reads from one replicate for each sample at the t=0 and t=24 h time points to 16S sequences assembled from our genomic data from the two plots from which the microcosm soil was obtained. A subset of 4,000 RNA reads was compared to the SILVA 16S database using BLAST to determine the percentage of RNA reads that were 16S rRNAs. Of 16S rRNA reads in the RNA data, $47\% \pm 19\%$ were determined to be at least 98% identical to 16S sequences assembled in the genomic data (Supplementary Table 12), which indicates that the



community that we assembled in the genomic dataset is a substantial fraction of the active community in the metatranscriptomic data.

We performed weighted gene co-expression network analyses using the WGCNA package³³ separately and individually on genes from seven genomes that were identified as having differentially expressed biosynthetic gene clusters over time, reasoning that these genomes will have the strongest signal of secondary metabolite co-expression. Transcripts per million for each gene were log-transformed. A soft network threshold was generated by choosing the lowest value that returned an \mathbb{R}^2 fit to a scale-free network greater than 0.8. A signed adjacency matrix was built using Pearson correlations, and a topographical overlap matrix was generated from the adjacency matrix. Module detection was run using the cuttreeDynamic() function with the 'hybrid' method, a minimum cluster size of 15, deepSplit set to TRUE and a cutHeight of 0.95.

Statistical analysis. To test whether cluster genes were significantly more coexpressed than random genes across a genome, we calculated all Spearman correlations between genes within clusters (mean $\rho\!=\!0.063; n\!=\!5,\!940$ comparisons), and compared this distribution of correlations to a distribution of all Spearman correlations between 100 randomly chosen genes from each genome (mean $\rho\!=\!0.041; n\!=\!503,\!699$ comparisons) using an independent two-group Wilcoxon rank-sum test ($P\!<\!0.001$). We also compared both distributions to a distribution of randomly selected genes from the entire dataset compared (mean $\rho\!=\!0.026$ $n\!=\!4947228$ comparisons) and found random genes to have the lowest levels of co-expression ($P\!<\!0.001$).

To identify differentially expressed clusters of genes between time points, we used the adonis function from the vegan package⁵². Transcript abundances in transcripts per million were log₂-transformed, and adonis tests were run on all clusters with any expression data for at least five proteins. *P* values were corrected for multiple tests using the Benjamini and Hochberg⁵³ method with a controlled family wise error rate of 5%.

To detect differential expression of individual genes within differentially expressed biosynthetic clusters between time points, we modelled Kallisto counts in the context of all metadata variables (plot, depth, treatment and time) using a negative binomial model implemented in DESeq2³². Kallisto count data from each genome were analysed independently so that the DEseq size factors for cross-sample count normalization would reflect the total transcriptomic activity of that genome in each sample. This approach is robust to biases in total transcriptomic activity per organism between samples, with the intention to identify differences in gene expression independent of changes in taxonomic composition, similar to previously reported methods³⁰. After size factor normalization, counts were fit to a negative binomial model of the form: count \sim depth + plot + treatment + time. To specifically test whether any genes exhibit differential expression associated with changes in time while accounting for the effects of depth, plot and treatment, we fit count data to a reduced model of the form: count \sim depth + plot + treatment. We then compared fits between the full and reduced model using the likelihood ratio test implemented in DESeq2. The significant genes (with an FDR-corrected P < 0.05) identified by comparing the full and reduced model were grouped, and direct comparisons were made between counts at 0 h and all other time points, to find those time points that exhibited a significant change in expression relative to the 0 h time point. This method confirmed differential expression of several individual genes within each differentially expressed biosynthetic cluster.

When examining modules of co-expression genes, the hypergeometric test was used to determine whether a module was significantly enriched in biosynthetic genes, using the phyper function in R.

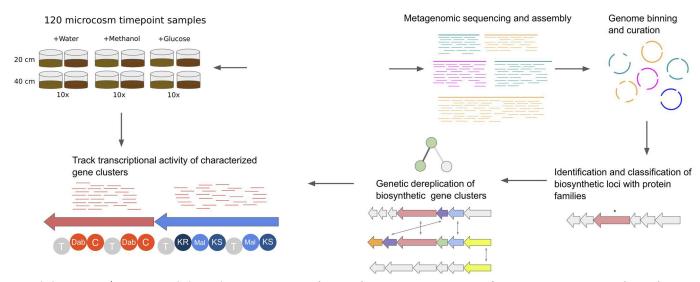
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Custom code used for the analyses (transcriptomics expression, DESeq2 differential expression and WGCNA co-expression analyses) that support this work is available in R Notebook format at http://www.github.com/alexcritschristoph/angelo_biosynthetic_genes_analysis.

Data availability. All genomic data associated with this project has been deposited in BioProject under accession PRJNA449266. DNA sequencing reads for this project have been deposited in the Sequence Read Archive database under PRJNA449266. Genomes analysed as part of this project have been submitted to the Whole Genome Shotgun (WGS) database. Genomes are also available through ggKBase at the following URL: http://ggkbase.berkeley.edu/angelo2014/organisms. Raw data for Fig. 2a and AntiSMASH annotated GenBank files for biosynthetic gene clusters reported on in this Letter are available at: http://www.github.com/alexcritschristoph/angelo_biosynthetic_genes_analysis.

- Bushnell, B. BBMap short read aligner. http://sourceforge.net/projects/bbmap (University of California, Berkeley, 2016).
- Joshi, N. A. & Fass, J. N. sickle a windowed adapative trimming tool for FastQ files (version 1.33) https://github.com/najoshi/sickle (2011).
- 38. Andrews, S. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
- Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a de novo assémbler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428 (2012).
- Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
- Brown, C.T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523, 208–211 (2015).
- Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607 (2016).
- Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144–1146 (2014).
- Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3, e1165 (2015).
- Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Methods* https://doi.org/10.1038/s41564-018-0171-1 (2018).
- Banfield, J. Development of a Knowledgebase to Integrate, Analyze, Distribute, and Visualize Microbial Community Systems Biology Data. Report No. DOE-UCB-4918) (US Department of Energy, 2015).
- Anantharaman, K. et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* 7, 13219 (2016).
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055 (2015).
- 49. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
- Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797 (2004).
- Price, M. N., Dehal, P. S. and Arkin, A. P. FastTree 2-approximately maximumlikelihood trees for large alignments. *PloS ONE* 5, e9490 (2010).
- 52. Oksanen, J. et al. vegan: Community ecology package https://cran.r-project.org/package=vegan (2007).
- Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57, 289–300 (1995).





Extended Data Fig. 1 | **Experimental plan and project overview.** Schematic showing major components of microcosm time-point sampling and metagenomic analyses.





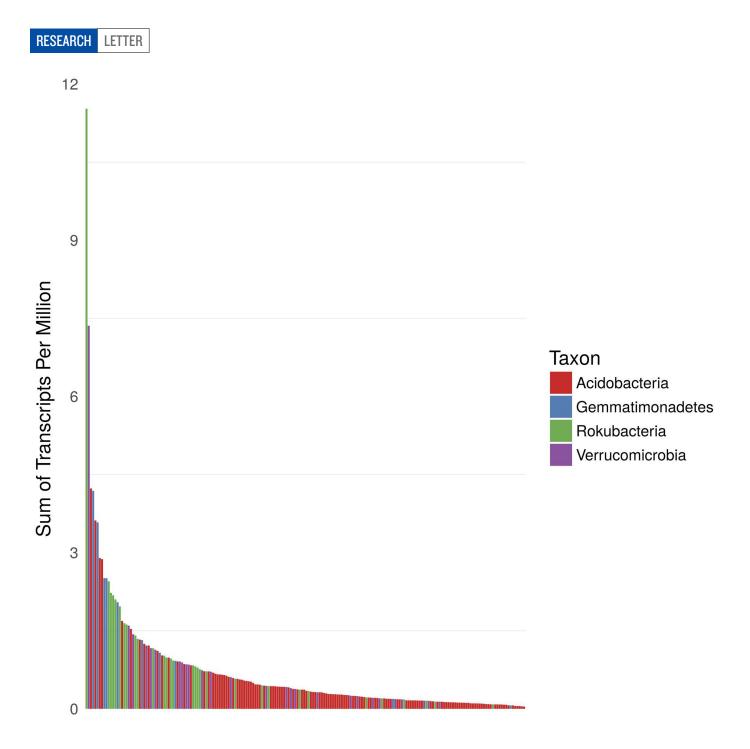
Extended Data Fig. 2 | **NRPS and PKS biosynthetic loci of the** *Candidatus* **Eelbacter genome.** Biosynthetic loci identified by both antiSMASH and PRISM from the *Candidatus* Eelbacter genome that contained at least 10 kb of biosynthetic genes. Predictions of the

organization of the biosynthetic domains in each locus shown here were determined by PRISM. Smaller biosynthetic loci from this genome are not shown. Full names for the biosynthetic domains are given in Supplementary Table 11.



Extended Data Fig. 3 | NRPS and PKS biosynthetic loci of the *Candidatus* Angelobacter genome. Biosynthetic loci identified by both antiSMASH and PRISM from the *Candidatus* Angelobacter genome that contained at least 10 kb of biosynthetic genes. Predictions of the

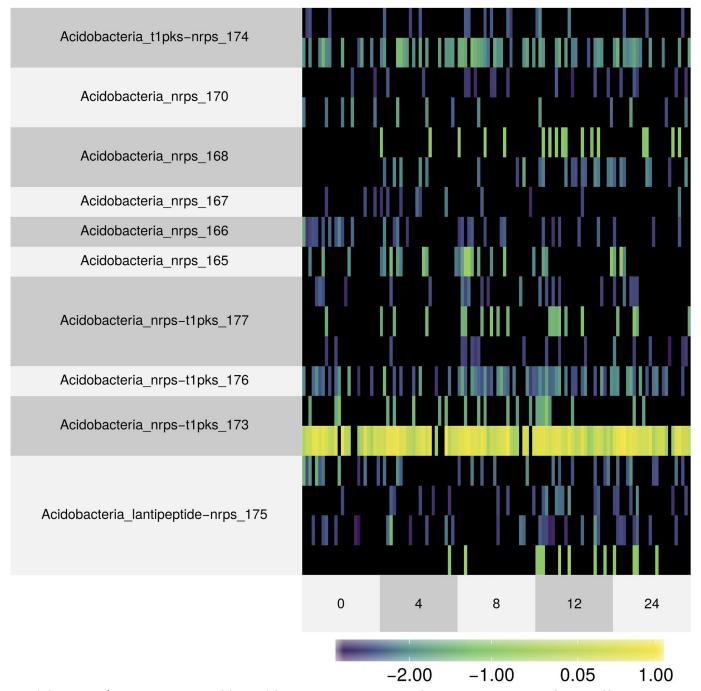
organization of the biosynthetic domains in each locus shown here were determined by PRISM. Smaller biosynthetic loci from this genome are not shown. Full names for the biosynthetic domains are given in Supplementary Table 11.



NRPS and PKS Proteins

Extended Data Fig. 4 | **Metatranscriptomics of NRPS and PKS proteins.** The graph shows levels of transcriptional expression of genes containing NRPS and PKS protein domains across genomes from the four phyla of

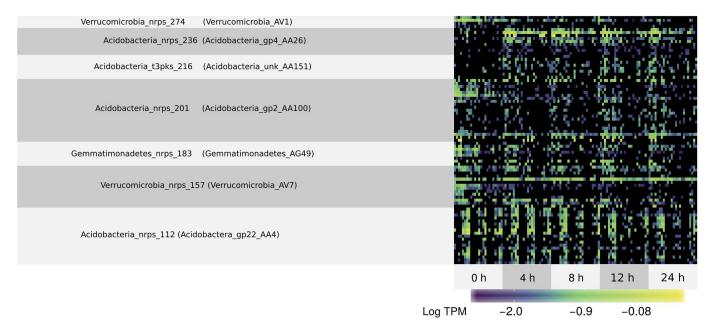
interest. Values are reported in \log_{10} -transformed transcripts per million and are summed across the 120 soil microcosm samples.



Extended Data Fig. 5 | Metatranscriptomics of the Candidatus Eelbacter genome. The levels of transcriptional expression of genes from biosynthetic gene clusters encoded in the Candidatus Eelbacter genome

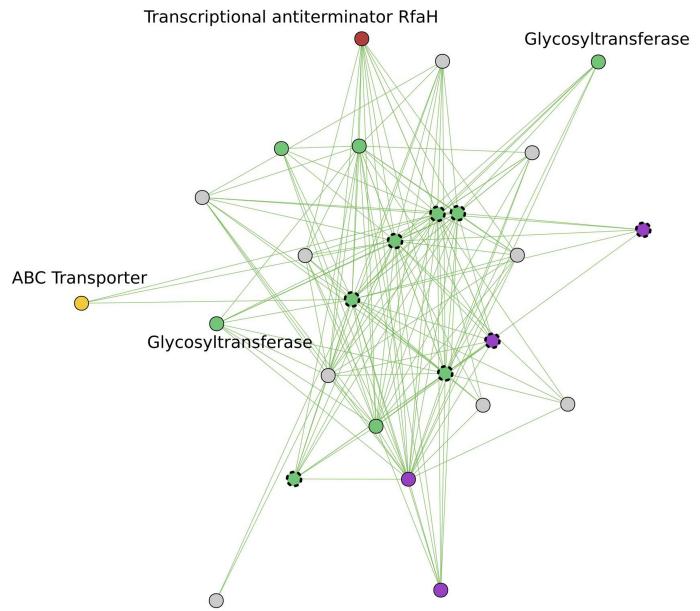
across 120 soil microcosm time-point samples grouped by extraction times (reported in hours) are shown. Expression levels are reported in \log_{10} -transformed transcripts per million.





Extended Data Fig. 6 | Differentially expressed biosynthetic gene clusters over time. The levels of expression of biosynthetic gene clusters from all organisms studied (excluding *Candidatus* Angelobacter data shown in Fig. 3a) that were found to be significantly differentially

expressed between time points (PERMANOVA; n=120; P<0.05, FDR = 5%) across 120 soil microcosm time-point samples are shown. Expression levels are reported in \log_{10} transcripts per million.



Extended Data Fig. 7 | Biosynthetic co-expression transcriptional module from Verrucomicrobia_AV7. A transcriptional network of co-expressed Verrucomicrobia_AV7 genes from a module found to be significantly enriched in genes from the biosynthetic gene clusters

Verrucomicrobia_nrps_156 and Verrucomicrobia_nrps_157 (P < 0.05; hypergeometric distribution) is shown. Genes from the biosynthetic locus are outlined with a dashed line.



Protection from UV light is an evolutionarily conserved feature of the haematopoietic niche

Friedrich G. Kapp^{1,2,3}, Julie R. Perlin^{1,2}, Elliott J. Hagedorn^{1,2}, John M. Gansner⁴, Daniel E. Schwarz⁵, Lauren A. O'Connell⁶, Nicholas S. Johnson⁷, Chris Amemiya⁸, David E. Fisher⁹, Ute Wölfle¹⁰, Eirini Trompouki¹¹, Charlotte M. Niemeyer³, Wolfgang Driever¹² & Leonard I. Zon^{1,2}*

Haematopoietic stem and progenitor cells (HSPCs) require a specific microenvironment, the haematopoietic niche, which regulates HSPC behaviour^{1,2}. The location of this niche varies across species, but the evolutionary pressures that drive HSPCs to different microenvironments remain unknown. The niche is located in the bone marrow in adult mammals, whereas it is found in other locations in non-mammalian vertebrates, for example, in the kidney marrow in teleost fish. Here we show that a melanocyte umbrella above the kidney marrow protects HSPCs against ultraviolet light in zebrafish. Because mutants that lack melanocytes have normal steady-state haematopoiesis under standard laboratory conditions, we hypothesized that melanocytes above the stem cell niche protect HSPCs against ultraviolet-light-induced DNA damage. Indeed, after ultraviolet-light irradiation, unpigmented larvae show higher levels of DNA damage in HSPCs, as indicated by staining of cyclobutane pyrimidine dimers and have reduced numbers of HSPCs, as shown by cmyb (also known as myb) expression. The umbrella of melanocytes associated with the haematopoietic niche is highly evolutionarily conserved in aquatic animals, including the sea lamprey, a basal vertebrate. During the transition from an aquatic to a terrestrial environment, HSPCs relocated into the bone marrow, which is protected from ultraviolet light by the cortical bone around the marrow. Our studies reveal that melanocytes above the haematopoietic niche protect HSPCs from ultraviolet-lightinduced DNA damage in aquatic vertebrates and suggest that during the transition to terrestrial life, ultraviolet light was an evolutionary pressure affecting the location of the haematopoietic niche.

Many aspects of the haematopoietic niche have been elucidated^{3,4}. However, little is known about the selective pressures during evolution that influenced the location of the niche in diverse tissues such as the bones in mammals and the kidney in teleost fish. One year after the hypothesis of the existence of a specialized niche for HSPCs in 1978⁵, it was hypothesized that HSPCs evolved to reside in the bone marrow of terrestrial animals to seek protection from ionizing irradiation, with bone then fulfilling the protective role of water⁶. Although this hypothesis is attractive, ionizing irradiation is mostly filtered out by Earth's atmosphere and there is no direct evidence that HSPCs would be susceptible to DNA damage by non-ionizing irradiation such as ultraviolet B (UVB) light in vivo and that this vulnerability could determine the location and characteristics of the haematopoietic niche.

To better understand the definitive haematopoietic niche in zebrafish, we examined HSPCs in their surrounding tissues, using the Tg(runx:mCherry) line that specifically labels HSPCs⁷. We were intrigued to find that an umbrella of internal melanocytes located dorsal

to the kidney marrow obscured visualization of HSPCs throughout development (Fig. 1, left, and Extended Data Fig. 1a, top left). HSPCs could more easily be observed in nacre mutants, which lack all melanocytes due to a mutation in the transcription factor *mitfa* (Fig. 1, right, and Extended Data Fig. 1a, bottom left; see also Extended Data Table 1). We confirmed the close spatial relationship between melanocytes and kidney HSPCs in larvae carrying Tg(mitfa:GFP) and Tg(runx:mCherry) transgenes that label melanocytes and HSPCs, respectively (Extended Data Fig. 1a, bottom right and Supplementary Video 1). To determine whether melanocytes serve as classical niche cells that support HSPC homing, expansion or maintenance, we compared HSPC numbers in mitfa^{-/-} larvae and their pigmented siblings at 5 and 7.5 days post fertilization by whole-mount in situ hybridization analysis of the expression of cmyb—a transcription factor and master regulator of vertebrate haematopoiesis. These time points assess the homing of HSPCs into the kidney marrow and the initial expansion therein⁸. Equivalent numbers of HSPCs that were positive for *cmyb* were present in pigmented and unpigmented siblings, and all larvae had the same staining intensity in the thymus, kidney and caudal haematopoietic tissue as a representatively stained pigmented larva (Extended Data Fig. 1b). We also evaluated adult haematopoiesis in different pigment-deficient fish by analysing the kidney marrow of 2-6-monthold Tg(runx:mCherry) casper mutant fish and their pigmented siblings (Extended Data Table 1) by flow cytometry. Neither HSPC abundance nor the relative abundance of blood progenitors, myelomonocytes or lymphocytes were affected by pigment loss (Extended Data Fig. 1c-f). In summary, we conclude that melanocytes are dispensable for steadystate haematopoiesis under laboratory lighting conditions.

Because the melanocytes form an opaque umbrella dorsal to the kidney marrow, we hypothesized that melanocytes shield HSPCs from UV-induced DNA damage, similar to their role in the skin. To test this, we irradiated unpigmented mitfa^{-/-} larvae and their pigmented siblings and assayed the most common form of UV-induced DNA damage, cyclobutane pyrimidine dimers (CPDs), in HSPCs. After UVC irradiation, kidney HSPCs positive for Tg(runx:mCherry) were immediately isolated by fluorescence-activated cell sorting (FACS) (Fig. 2a) and stained with an antibody that recognizes CPDs. After irradiation HSPCs isolated from unpigmented larvae showed higher levels of DNA damage than HSPCs isolated from pigmented larvae (Extended Data Fig. 2a); this difference was significant when the fluorescence per cell was quantified ($P \le 0.01$, Fig. 2b). Since UVC light is usually filtered out by Earth's atmosphere, we next tested biologically more relevant light (UVB), which also penetrates well into clear water. We exposed larvae to UVB and performed paraffin sections afterwards to assess

¹Department of Stem Cell and Regenerative Biology and Harvard Stem Cell Institute, Harvard University, Cambridge, MA, USA. ²Stem Cell Program and Division of Hematology/Oncology, Boston Children's Hospital and Dana Farber Cancer Institute, Howard Hughes Medical Institute, Harvard Stem Cell Institute, Harvard Medical School, Boston, MA, USA. ³Department of Pediatric Hematology and Oncology, Center for Pediatrics, Medical Center–University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany. ⁴Division of Hematology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. ⁵US Fish and Wildlife Service, Private John Allen National Fish Hatchery, Tupelo, MS, USA. ⁶Department of Biology, Stanford University, Stanford, CA, USA. ⁷US Geological Survey, Great Lakes Science Center, Hammond Bay Biological Station, Millersburg, MI, USA. ⁸Molecular Cell Biology, University of California, Merced, CA, USA. ⁹Cutaneous Biology Research Center, Massachusetts General Hospital and Harvard Medical School, Charlestown, MA, USA. ¹⁰Department of Dermatology, Medical Center–University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany. ¹¹Department of Cellular and Molecular Immunology, Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany. ¹²Developmental Biology, Faculty of Biology, Centre for Biological Signalling Studies (BIOSS), Albert-Ludwigs-University of Freiburg, Freiburg, Germany. *e-mail: zon@enders.tch.harvard.edu

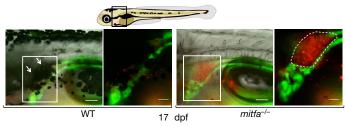


Fig. 1 | **Melanocytes are spatially associated with the zebrafish kidney marrow.** Zebrafish positive for Tg(cdh17:GFP) (green, labelling the kidney tubule) and Tg(runx:mcherry) (red, labelling HSPCs) are depicted at 17 days post-fertilization (dpf). The boxed area in the schematic of the embryo is enlarged in the bright field panels below. The boxed areas in the bright field images are enlarged in the fluorescence images on the right. Fluorescence images show the head kidney containing the haematopoietic marrow (indicated by the dashed outline). The white arrows highlight the melanocyte umbrella. Scale bars, 100 μm (bright field) and 50 μm (fluorescence). WT, wild type.

the spatial association of melanocytes and CPD⁺ cells. Only very few CPDs were present directly below the melanocytes compared to the rest of the larva (Fig. 2c) and the thymus, an organ also protected by melanocytes, only showed CPD⁺ cells in unpigmented larvae (Fig. 2d and Extended Data Fig. 2c), confirming the known protective role of melanocytes against UV light⁹.

On the basis of the observation that HSPCs accumulate UV-induced DNA damage in unpigmented larvae, we evaluated the functional effects of UV irradiation on HSPCs in the presence or absence of the melanocyte umbrella by using *mitfa*^{-/-} mutants and their pigmented siblings (Fig. 3a, b). Two days after UVB irradiation, unpigmented mitfa^{-/-} larvae showed a significant decrease in HSPC numbers as assessed by *cmyb* staining compared to their non-irradiated siblings (P = 0.001), whereas their pigmented siblings did not have a significant decrease in *cmyb* staining (Fig. 3c). This was also true in sandy $(tyr^{-/-})$ mutants and larvae treated with 1-phenyl 2-thiourea (PTU) that each have normal melanocyte numbers but cannot synthesize melanin (Fig. 3c and Extended Data Table 1; P = 0.008 and P = 0.016, respectively). We could replicate this effect with the more damaging UVC in PTU-treated larvae as well as in $tyr^{-/-}$ mutants irradiated with the same UVB dose at a lower UV index for a longer period of time (Extended Data Fig. 3a-c). These results indicate that melanocytes containing melanin are required to prevent the detrimental effects on HSPCs caused by exposure to UV in zebrafish larvae.

We next assessed whether the orientation of the melanocyte umbrella was important for its protective function or whether a mechanism independent of optical shielding, such as paracrine signalling, could be involved. Larvae were anaesthetized, causing them to turn on their back, thus moving the otherwise unperturbed umbrella of pigmented melanocytes out of the path of light and exposing the HSPCs to UV irradiation from the ventral side (Fig. 3d). The *cmyb* staining in pigmented, anaesthetized larvae was reduced to the same level as in unpigmented, non-anaesthetized larvae after UVB treatment (Fig. 3e), and both were significantly lower than baseline (P=0.024 and P=0.014, respectively, compared to non-pigmented, non-irradiated, anaesthetized larvae). This shows that the orientation of the melanocyte umbrella is critical to prevent damage to HSPCs by UV light and suggests that melanocytes protect HSPCs through a purely optical shielding mechanism.

To confirm the adverse effect of UV light on an unprotected haematopoietic system, we examined the cellularity of the kidney marrow and found that it was markedly decreased in non-pigmented, irradiated larvae compared to pigmented, irradiated siblings and non-irradiated controls (Extended Data Fig. 3d). In addition, the numbers of circulating thrombocytes positive for Tg(CD41:GFP) were significantly lower in non-pigmented, irradiated larvae than in the other groups (Extended Data Fig. 3e, f), suggesting a reduced output of differentiated blood cells from unprotected HSPCs.

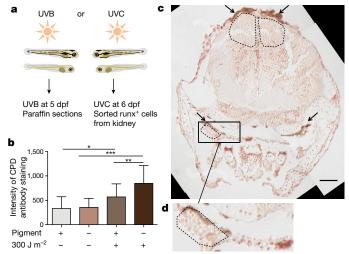


Fig. 2 | Melanocytes protect HSPCs from UV-induced DNA damage. a, Experimental layout. Animals were exposed to UV light at either 5 or 6 dpf and were processed immediately after UV irradiation. b, Quantification of immunostaining intensity per cell. Data are mean \pm s.d., data were analysed by ANOVA with post hoc Bonferroni's multiple comparison test. HSPCs from 5 kidneys for each condition; n=45, 33, 16 and 33 isolated cells from left to right. Dot plots are shown in Extended Data Fig. 2b. c, Anti-CPD immunostaining of a pigmented irradiated fish (transverse paraffin section). Black arrows indicate melanocytes. Dashed black lines indicate areas below the melanocytes with only few CPD+ nuclei. Scale bar, 50 μ m. d, Magnification of the indicated area (thymus) in c. *P < 0.05; *P < 0.01; **P < 0.0001.

The observed adverse effect of UV light on HSPCs is consistent with previous in vitro studies, in which it was shown that much lower UVB doses of 100–200 J m⁻² completely abrogated colony forming potential of human HSPCs¹⁰. We chose a UVB dose that corresponds to a sunlight exposure of approximately 10-20 min at UV indices (a measurement of sunlight intensity) of 5–10; this dose corresponds to a UV exposure that would give a fair-skinned person sunburn. Wild zebrafish are found in rice paddies and small, often clear pools¹¹, and other fish species also swim close to the water surface during the middle of the day in clear water (Extended Data Fig. 4). Since UVB penetrates well into clear water¹², HSPCs in fish would be exposed to UV light in natural conditions and would thus benefit from optical protection. Fish have evolved other protective mechanisms against the accumulation of UV-induced DNA damage, such as light-dependent photoenzymatic repair¹³ and the expression of a sunscreen compound, gadusol¹⁴. These findings highlight the importance of coping strategies against UV-induced DNA damage even in aquatic animals.

To test whether the protective melanocyte umbrella was specific to zebrafish larvae, we performed comparative histology along the evolutionary tree of fish and other vertebrates (Fig. 4a). The adult zebrafish kidney is covered with melanocytes, which can readily be identified on histology slides stained with haematoxylin and eosin (Fig. 4f). We found that all analysed teleost fish (Ictalurus punctatus (channel catfish), Gasterosteus aculeatus (stickleback), Lepomis macrochirus (blue gill) and Lepomis microlophus (redear sunfish), Fig. 4e, g-i) had melanocytes covering the haematopoietic kidney marrow. This was also true in a member of the chondrostei, Acipenser fulvescens (lake sturgeon), and a member of the holostei, Atractosteus spatula (alligator gar) (Fig. 4c, d), that both diverged from teleost species about 250–350 million years ago^{15,16}. Even the ancestral jawless vertebrate Petromyzon marinus, at the base of the vertebrate lineage (sea lamprey, post metamorphic stage), which diverged approximately 500 million years ago, showed melanocytes around its haematopoietic niche in the supraspinal organ¹⁷ (Fig. 4b). In more-recently diverged aquatic members of the vertebrate lineage, such as the sarcopterygian Protopterus annectens (West African lungfish), melanocytes also covered the

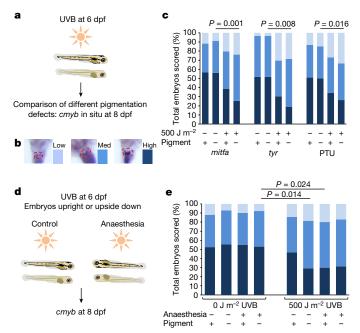


Fig. 3 | UV exposure is detrimental to unprotected HSPCs. a, Experimental layout. Animals were exposed to UV light at 6 dpf and fixed at 8 dpf. b, Scoring scheme for *cmyb* in situ (RNA) in the zebrafish kidney (dashed red outline). The colour of the scoring scheme corresponds to the colour of the bars in c and e. c, Relative distribution of HSPC abundance in various pigment-deficient larvae. Lack of mitfa leads to the absence of melanocytes, whereas $tyr^{-/-}$ mutants and PTU-treated larvae do not produce melanin. 'Pigment' indicates the pigmentation phenotype and is a result of a mutation in *mitfa* or tyr ($mitfa^{-/-}$ or $tyr^{-/-}$ mutants have no pigmentation, indicated with minuses in 'Pigment') or treatment with PTU (PTU treatment leads to no pigmentation, indicated with a minus in 'Pigment'. n = 51, 57, 49, 67, 31, 31, 23, 21, 53, 54, 67 and 57 larvae from left to right. d, Experimental layout for experiment that included anaesthesia. e, Relative distribution of HSPC abundance in the different treatment groups. n = 59, 56, 51, 64, 49, 65, 50, and 35 larvae in from left to right. Data were analysed by χ^2 test.

kidney marrow (Fig. 4j). In aquatic tetrapod larvae, from the amphibians Xenopus laevis (African clawed frog) and Epipedobates anthonyi (Anthony's poison arrow frog), a melanocyte umbrella was present above the haematopoietic niche in the liver and kidney, respectively (Fig. 4k, 1). It is known that in terrestrial amphibians, the adult haematopoietic niche is located in the bone marrow, which we confirmed in *Phyllobates terribilis* (golden poison frog; Fig. 4m). Using the anuran amphibian Dendrobates tinctorius (Dyeing poison frog; Fig. 4n) and analysing its haematopoietic niche at different developmental time points from tadpole to froglet (Extended Data Fig. 5a), we were able to show that the transition from a melanocyte-covered niche to the bone marrow occurred when the tadpoles first develop legs while still in an aquatic environment (Extended Data Fig. 5b-f). We then confirmed that cortical bone does indeed provide shielding against UV light by performing anti-CPD immunostaining of a hind leg of the *D*. tinctorius tadpole depicted in Extended Data Fig. 5e that was exposed to UVB light post mortem (Extended Data Fig. 6). This finding might indicate that the cortical bone around the bone marrow serves as a UV-protective layer in lieu of melanocytes, which might explain why all terrestrial animals have their haematopoietic niche in the bone marrow. Notably, some mammals have genetically programmed melanocytes in the spleen¹⁸, and these melanocytes might represent an evolutionary remnant of the melanocyte umbrella that we discovered in zebrafish. Some frog species, such as certain Rana species, exhibit shifting sites of haematopoiesis in adulthood with the bone marrow serving as the main and the liver as a minor haematopoietic site 19,20. In addition, seasonal variation can be observed in these species, with the liver being more haematopoietic during the winter and the bone

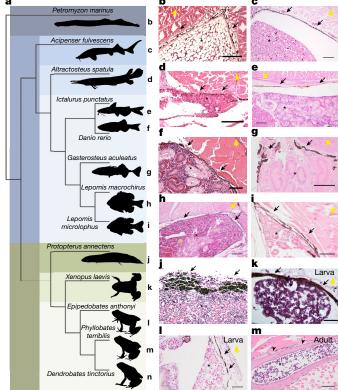


Fig. 4 | Melanocytes are present around the haematopoietic niche of aquatic animals. a, Phylogenetic tree based on cytochrome oxidase I sequences. Species-specific histology analyses are shown in panels \mathbf{b} — \mathbf{m} as indicated. \mathbf{b} — \mathbf{m} , Haematoxylin and eosin staining of haematopoietic tissues. Kidneys were identified by the presence of tubules; haematopoietic activity was inferred by abundance of haematopoietic cells. Black arrows indicate the melanocyte layer; arrowheads indicate the cortical bone; asterisks indicate the kidney or bone marrow; yellow arrows indicate the orientation of the animal (kidney sections only; yellow tip towards the dorsal aspect). Scale bars, $50 \ \mu \mathbf{m}$ (\mathbf{f} , \mathbf{g} , \mathbf{i} , \mathbf{l}), $100 \ \mu \mathbf{m}$ in (\mathbf{b} – \mathbf{e} , \mathbf{h} , \mathbf{m}) and $200 \ \mu \mathbf{m}$ (\mathbf{j}).

marrow during the summer²¹, which might reflect an adaption to changing UV levels.

We hypothesize that during the evolution of tetrapods, UV light was a selective pressure in for the location of the HSPC niche. Larvae in which HSPCs colonized the bone marrow before the transition from aquatic to terrestrial life were at an advantage due to the selective pressures from higher UV levels in terrestrial conditions, although other factors, such as a more hypoxic microenvironment, might also have contributed to this process. Our hypothesis is also consistent with the development of the bone marrow in early sarcopterygians (lobe-finned fish)²² and with the observation that traits of terrestrial animals were often acquired before the transition out of the water^{23–25}. Our studies provide evidence for the hypothesis stated by Edwin Cooper in 1979⁶ as to why HSPCs are located in the bone marrow of terrestrial animals, where they find shelter from harmful irradiation.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0213-0.

Received: 3 July 2017; Accepted: 15 May 2018; Published online 13 June 2018.

- 1. Ding, L. & Morrison, S. J. Haematopoietic stem cells and early lymphoid
- progenitors occupy distinct bone marrow niches. Nature 495, 231–235 (2013).
 Kunisaki, Y. et al. Arteriolar niches maintain haematopoietic stem cell quiescence. Nature 502, 637–643 (2013).



- Acar, M. et al. Deep imaging of bone marrow shows non-dividing stem cells are mainly perisinusoidal. Nature 526, 126–130 (2015).
- Chen, J. Y. et al. Hoxb5 marks long-term haematopoietic stem cells and reveals a homogenous perivascular niche. Nature 530, 223–227 (2016).
- Schofield, R. The relationship between the spleen colony-forming cell and the haemopoietic stem cell. Blood Cells 4, 7–25 (1978).
- Horton, J. D. Development and differentiation of vertebrate lymphocytes: review of the Durham symposium — September 1979. Dev. Comp. Immunol. 4, 177–181 (1980).
- Tamplin, O. J. et al. Hematopoietic stem cell arrival triggers dynamic remodeling of the perivascular niche. Cell 160, 241–252 (2015).
- Murayama, E. et al. Tracing hematopoietic precursor migration to successive hematopoietic organs during zebrafish development. *Immunity* 25, 963–975 (2006)
- Kaidbey, K. H., Agin, P. P., Sayre, R. M. & Kligman, A. M. Photoprotection by melanin—a comparison of black and Caucasian skin. J. Am. Acad. Dermatol. 1, 249–260 (1979).
- Azuma, H. et al. Comparison of sensitivity to ultraviolet B irradiation between human lymphocytes and hematopoietic stem cells. *Blood* 96, 2632–2634 (2000).
- Èngeszer, R. E., Patterson, L. B., Rao, A. A. & Parichy, D. M. Zebrafish in the wild: a review of natural history and new notes from the field. *Zebrafish* 4, 21–40 (2007).
- Tedetti, M. et al. High penetration of ultraviolet radiation in the south east Pacific waters. Geophys. Res. Lett. 34, L12610 (2007).
- Mitchell, D. L., Meador, J. A., Byrom, M. & Walter, R. B. Resolution of UVinduced DNA damage in *Xiphophorus* fishes. *Mar. Biotechnol.* 3, S61–S71 (2001).
- Osborn, A. R. et al. De novo synthesis of a sunscreen compound in vertebrates. eLife 4, e05919 (2015).
- Volff, J.-N. Genome evolution and biodiversity in teleost fish. Heredity 94, 280–294 (2005).
- 16. Hurley, I. A. et al. A new time-scale for ray-finned fish evolution. *Proc. R. Soc. B* **274**, 489–498 (2007).
- Amemiya, C. T., Saha, N. R. & Zapata, A. Evolution and development of immunological structures in the lamprey. *Curr. Opin. Immunol.* 19, 535–541 (2007).
- Weissman, I. Genetic and histochemical studies on mouse spleen black spots. Nature 215, 315 (1967).
- de Abreu Manso, P. P., de Brito-Gitirana, L. & Pelajo-Machado, M. Localization of hematopoietic cells in the bullfrog (*Lithobates catesbeianus*). Cell Tissue Res. 337, 301–312 (2009).
- Akiyoshi, H. & İnoue, A. M. Comparative histological study of hepatic architecture in the three orders amphibian livers. *Comp. Hepatol.* 11, 2 (2012).
- Maslova, M. N. & Tavrovskaia, T. V. [The seasonal dynamics of erythropoiesis in the frog *Rana temporaria*]. *Zh. Evol. Biokhim. Fiziol.* 29, 211–214 (1993).
 Sanchez, S., Tafforeau, P. & Ahlberg, P. E. The humerus of *Eusthenopteron*: a
- Sanchez, S., Tafforeau, P. & Ahlberg, P. E. The humerus of Eusthenopteron: a puzzling organization presaging the establishment of tetrapod limb bone marrow. Proc. R. Soc. B 281, 20140299 (2014).
- Niedźwiedzki, G., Szrek, P., Narkiewicz, K., Narkiewicz, M. & Ahlberg, P. E. Tetrapod trackways from the early Middle Devonian period of Poland. *Nature* 463, 43–48 (2010).
- Markey, M. J. & Marshall, C. R. Terrestrial-style feeding in a very early aquatic tetrapod is supported by evidence from experimental analysis of suture morphology. *Proc. Natl Acad. Sci. USA* **104**, 7134–7138 (2007).
 MacIver, M. A., Schmitz, L., Mugan, U., Murphey, T. D. & Mobley, C. D. Massive
- MacIver, M. A., Schmitz, L., Mugan, U., Murphey, T. D. & Mobley, C. D. Massive increase in visual range preceded the origin of terrestrial vertebrates. *Proc. Natl Acad. Sci. USA* 114, E2375–E2384 (2017).

Acknowledgements We thank D. Richardson at the Harvard Center for Biological Imaging for infrastructure and support; C. MacGillivray at the $\ensuremath{\mathsf{HSCRB}}$ Histology Core and Joyce LaVecchio at the HSCRB Flow Cytometry Core for technical assistance; E. van Italie in M. Kirschner's laboratory and J. Cech in C. Peichel's laboratory for providing Xenopus and stickleback samples; the Zebrafish Atlas (http://zfatlas.psu.edu/, NIH grant R24 RR017441, Jake Gittlen Cancer Research Foundation, and PA Tobacco Settlement Fund) for provision of the adult zebrafish histology image. Any use of trade, product or firm names is for descriptive purposes only and does not imply endorsement by the US Government. This work was supported by HHMI and NIH grants 5P01 CA163222, R01 HL048801, P01 HL032262, U54 DK110805-01, R01 DK053298, U01 HL100001-05, and R24 DK092760 to L.I.Z. D.E.F. acknowledges grant support from NIH (5P01 CA163222 and 2R01 AR043369) and the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation. E.J.H. was supported by 1K01DK111790-01. Further support came from the German Research Foundation (DFG-SFB850-A1, to W.D.) and the Excellence Initiative of the German Federal and State Governments (Centre for Biological Signalling Studies EXC 294, to W.D.). F.G.K. was supported by a postdoctoral fellowship of the German Cancer Aid (70110820), a return scholarship of the Forschungskommission, Faculty of Medicine, University of Freiburg, and an EXCEL-Fellowship of the Faculty of Medicine, University of Freiburg, funded by the Else-Kröner-Fresenius-Stiftung. N.S.J. was supported by the Great Lakes Fishery Commission. E.T. was supported by funding from Max Planck Gesellschaft and a Marie Curie Career Integration grant (631432), a Fritz Thyssen Stiftung and the DFG founded Research Training Group GRK2344 'MelnBio – BiolnMe'. L.A.O. was supported by a Bauer Fellowship from Harvard University. J.M.G. was supported by T32 training grants (T32CA009172-39 and T32HL116324-03).

Reviewer information *Nature* thanks I. Beerman, G. Litman and S. Morrison for their contribution to the peer review of this work.

Author contributions F.G.K. planned, executed or analysed all experiments. F.G.K. and L.I.Z. wrote the manuscript with input from all authors. J.R.P. and E.T. helped with a subset of in situ experiments, J.M.G. with flow cytometry, E.J.H. with a subset of imaging experiments and L.A.O. created the phylogenetic tree as well as the animal drawings. D.E.S., L.A.O., N.S.J., C.A. and U.W. provided essential materials and samples for the manuscript. W.D., C.M.N., D.E.F. and L.I.Z. supervised the project and gave input to experimental design. All authors discussed the results and commented on the manuscript.

Competing interests L.I.Z. is a founder and stockholder of Fate Therapeutics, Inc., Scholar Rock and Camp4 Therapeutics. D.E.F. has a financial interest in Soltego, Inc., a company developing SIK inhibitors for topical skin darkening treatments that might be used for a broad set of human applications. The interests of D.E.F. were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0213-0.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0213-0.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.I.Z. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Zebrafish husbandry. Zebrafish maintenance and breeding was performed at 28.5 °C with a 14 h:10 h light:dark cycle²⁶. These standard laboratory conditions do not comprise exposure to UV light. All experiments were performed according to protocols approved by the Institutional Animal Care and Use Committees (IACUCs) of Harvard University and Boston Children's Hospital, or by the Regierungspräsidium Freiburg, and were in accordance with the German laws for animal care.

Frog methods. *E. anthonyi*, *P. terribilis* and *D. tinctorius* were reared in a captive colony. Animals were anaesthetized with 20% benzocaine followed by euthanasia by cervical transection. Specimens were then placed in 4% paraformaldehyde. All poison frog protocols were approved by the IACUC at Harvard University.

Statistics and reproducibility. Owing to animal welfare regulations in Germany, complete experiments involving zebrafish older than 5 dpf could only be performed once, since repeat experiments are not permissible once a statistically significant result has been obtained. Small-scale pre-experiments have been performed to estimate the effect strength in the assays performed, and the animal experiments were statistically planned and approved by the Regierungspräsidium Freiburg with sufficient numbers of animals to obtain statistically significant results. Confidence in the observed results and their reproducibility is strengthened by an experimental strategy, in which successive experiments not only investigate new biological questions, but are also based on and thus add supportive information to the preceding experiment (for example, irradiation of different pigment mutants in Fig. 3c, anaesthesia experiment in Fig. 3e, or thrombocyte count in Extended Data Fig. 3e). The sex of the animals was not or could not be determined before the conduction of the experiments.

The experiments in the figures were performed as follows: Figure 1: imaging experiments were repeated >3 times independently with similar results (biological replicates). Figure 2b: the experiment was performed once. Figure 2c: the experiment was performed twice with similar results (biological replicate). Figure 3c, e: the experiments were performed once. Figure 4b-m: experimental results were confirmed in at least a second animal of the same species (except lungfish, owing to scarcity of material). Extended Data Fig. 1a (and Supplementary Video 1): the experiment was performed at least twice with similar results (biological replicate). Extended Data Fig. 1b: the experiment was performed three times with similar results (biological replicates). Extended Data Fig. 1c-f: the experiment was performed twice with similar results (biological replicates). Extended Data Fig. 2a, b: the experiment was performed once (same experiment as Fig. 2b). Extended Data Fig. 2c: the experiment was performed twice with similar results (biological replicate). Extended Data Fig. 3b: The experiment was performed twice with similar results (biological replicate). Extended Data Fig. 3c: the experiment was performed once. Extended Data Fig. 3d-f: the experiments were each performed once (the same larvae were used for the thrombocyte count as in the histology experiments to reduce the number of animals used). Extended Data Fig. 4: N/A. Extended Data Fig. 5: the experiment was performed once owing to scarcity of material, but the analyses at different developmental time points support each other. Extended Data Fig. 6: the experiment was performed once owing to scarcity of material. Extended Data Fig. 7: N/A.

Experiments were conducted in a blinded fashion, whenever possible. Animals were randomly assigned to treatment and control groups.

Imaging. Using the recently developed transgenic reporter line $Tg(Mmu.Runx1:NLS-mCherry)^{cz2010}$ (here called Tg(runx:mCherry)) that labels HSPCs⁵, we imaged the location of HSPCs relative to the kidney tubule in a cross with the $Tg(cdh17:GFP)^{nz1}$ line²⁷ in zebrafish larvae at different developmental stages and also assessed the spatial relationship with melanocytes, labelled by the transgenic reporter line²⁸ $Tg(mitfa:GFP)^{w47}$. Fish were anaesthetized with 0.168 mg tricaine per ml egg water for the duration of the procedure and were imaged on a Zeiss CellObserver, Zeiss Examiner or a Zeiss LSM700 system. For the thrombocyte count, $Tg(CD41:GFP)^{la2}$ larvae²⁹ were analysed two days after irradiation by imaging on a Zeiss Examiner with a $20 \times$ objective and time lapse videos of 10 s were recorded. Afterwards, a z projection was performed and the circulating thrombocytes were counted and normalized to the area of the vessel (Extended Data Fig. 3f). Statistical analysis was performed using ANOVA with post hoc Bonferroni test. Image analysis and processing was performed with Image], ZEN (Zeiss) and PowerPoint (Microsoft).

Flow cytometry. Kidney marrow was isolated from adult wild-type, $mitfa^{-/-}$, $roy^{-/-}$ and $casper^{-/-}$ fish and analysed as previously described 30 . For gating strategy see Supplementary Fig. 1.

UVC and UVB irradiation. Pigmented and unpigmented larvae were placed in 6-cm Petri dishes and egg water was added to a volume of 10 ml. For each irradiation dose, pigmented and unpigmented larvae were placed in the same dish to achieve identical UV exposure in pigmented and unpigmented fish. For UVC irradiation a Stratalinker 1800 (Stratagene) was used, for UVB irradiation an UV 801 BL unit (Waldmann GmbH, Germany) was used. The Petri dish was placed

in a cardboard container to reduce the amount of UV light reaching the larvae from the side (see experimental setup displayed in Extended Data Fig. 7). The administered dose of UVB corresponded to approximately 10 min at a UV index of 10; both a UV index of 90 for approximately 50 s (for example, Fig. 3c, e) and a UV index of 20 for approximately 5 min (Extended Data Fig. 3c) led to the same results; of note, even a UV index of 20 is slightly higher than usually encountered in natural conditions. The UV index was measured using a Solarmeter 6.5 (Solar Light Company, Inc.). The dose was also measured with UV-Sensor Variocontrol (Waldmann, Germany) and corresponded to 500 J m $^{-2}$. Before and after the irradiation, larvae were kept in the dark at 28.5 °C until the end of the experiment. After irradiation, larvae were transferred to a 10-cm dish containing 30–40 ml of egg water to maintain good water quality.

PTU treatment. To prevent pigmentation in wild-type TU (Tübingen) embryos, embryos were treated with 150 μM 1-phenyl 2-thiourea (PTU; Sigma-Aldrich) at 24 hours post-fertilization (hpf), a slightly reduced dose of PTU that consistently gave very good results 31 . To avoid interaction of PTU with the UV light, PTU was washed off and the embryo medium was replaced 12 h before irradiation.

Anaesthesia. To move the melanocytes out of the path of light, larvae were anaesthetized immediately before the irradiation. Tricaine was added to the 6-cm Petri dish at a final concentration of 0.168 mg tricaine per ml egg water. After irradiation, tricaine was immediately washed off and replaced with egg water.

FACS and anti-CPD antibody staining. Staining of single cells positive for Tg(runx:mCherry) for UV-induced DNA damage was performed according to a protocol adapted from previously published methods³². Immediately after UV irradiation, larvae were euthanized in tricaine and transferred to ice water. The tails of the larvae were removed behind the swim bladder and discarded, since only HSPCs in the kidney were of interest. Five larval heads were pooled according to their pigment status, incubated with liberase at 37 °C for 20 min in the dark. FBS was added to reach 10% to inhibit further digestion and larvae were dissociated by pipetting up and down. Debris was removed by pipetting through a 40- μm mesh filter. Then, 10% formaldehyde was added to reach a 4% final concentration and cells were fixed for 10 min at room temperature. Afterwards, cells were spun down at 500g for 5 min at 4 °C and washed with PBS twice. Cells were sorted onto SuperFrost slides with attached 8-well silicone insulators containing $70-100~\mu l$ ultrapure water using a FACSAria (HSCRB FACS Core) with the 355-nm laser turned off. Slides were dried at room temperature, then put in an oven at approximately 70 °C for 10 min. Slides were kept in a humidified chamber for the following steps. Cells were washed with PBS, incubated with PBS with 0.5% Triton X-100 $\,$ for 20 min, washed with PBS, DNA was denatured with 2 M HCl for 30 min at room temperature, washed with PBS, blocked with 10% NGS for 60 min, and incubated with the anti-CPD antibody TDM-2 (Cosmo Bio Co) at 1:500 in 2% NGS overnight at 4°C. The next day, cells were washed with PBS and kept in the dark for the subsequent steps. The slides were incubated with a goat anti-mouse antibody linked to AlexaFluor 488 (Invitrogen) at 1:500 in 2% NGS for 30 min at room temperature. Cells were washed with PBS, followed by incubation with $0.05\,\mu g\,ml^{-1}$ DAPI in PBS for 5 min and washed again with PBS. Slides were analysed on a CellObserver (Zeiss, Germany) and the mean fluorescence of each cell was measured. Statistical analysis was performed using GraphPad Prism 5 software using a one-way ANOVA with a post hoc Bonferroni's multiple comparison test. Paraffin sections. Directly after irradiation, larvae were euthanized in Tricaine

followed by fixation in 4% formaldehyde. Larvae were euthanized in Iricane followed by fixation in 4% formaldehyde. Larvae were embedded in paraffin wax, sectioned at 5–7 μ m and immunohistochemistry was performed as described above with an additional proteinase K digestion step. The secondary biotinylated anti-mouse (Vector Laboratory) (1:200) was used followed by signal detection with VECTASTAIN ABC HRP kit (Vector Laboratory).

Whole-mount in situ hybridization. Whole-mount in situ hybridization was performed as previously described 33 with the addition of 0.2% glutaraldehyde to the 4% fixation step after permeabilization with proteinase K. The cmyb probe was used at 400 ng ml $^{-1}$. Staining with BCIP and NBT took approximately 4–6 h. Statistical analysis was performed using GraphPad Prism 5 software using χ^2 tests.

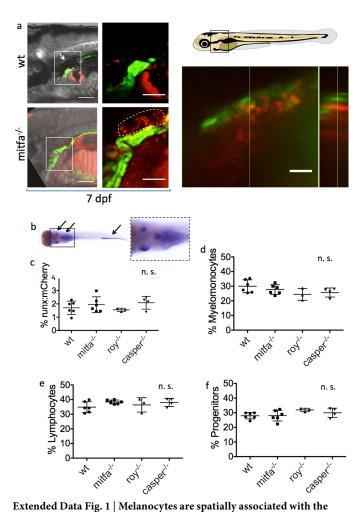
Histology. Fish specimens were euthanized and fixed in 4% formaldehyde for at least 24 hpf at 4 °C. After fixation, samples were transferred to 70% ethanol and stored at 4 °C until further processing. Samples were decalcified and afterwards embedded in paraffin wax. After hardening, samples were cut at a thickness of 4–10 μ m on a microtome, transferred to charged glass slides and stained with haematoxylin and eosin and covered with a coverslip afterwards. The zebrafish histology slide (Fig. 4f) was acquired from the Zebrafish Histology Atlas (http://bio-atlas.psu.edu/zf/view.php?s=250&atlas=18&z=1&c=9774,7525).

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request. Source data for Figs. 2, 3 and Extended Data Figs. 1–3, 6 are provided in the online version of the paper.



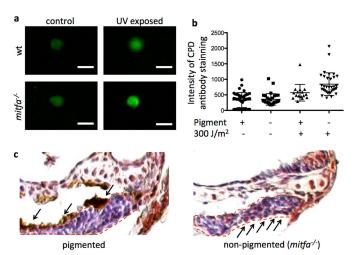
- 26. Westerfield, M. The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Danio rerio) 4th edn (University of Oregon Press, Eugene, 2000).
- Diep, C. Q. et al. Identification of adult nephron progenitors capable of kidney regeneration in zebrafish. *Nature* 470, 95–100 (2011).
- Curran, K., Raible, D. W. & Lister, J. A. Foxd3 controls melanophore specification in the zebrafish neural crest by regulation of Mitf. Dev. Biol. 332, 408–417 (2009).
- 29. Lin, H.-F. et al. Analysis of thrombocyte development in CD41–GFP transgenic zebrafish. *Blood* **106**, 3803–3810 (2005).
- 30. Traver, D. et al. Transplantation and in vivo imaging of multilineage engraftment in zebrafish bloodless mutants. *Nat. Immunol.* **4**, 1238–1246 (2003).
- Karlsson, J., von Hofsten, J. & Olsson, P.-E. Generating transparent zebrafish: a refined method to improve detection of gene expression during embryonic development. *Mar. Biotechnol.* 3, 522–527 (2001).
- 32. Èma, H. et al. Adult mouse hematopoietic stem cells: purification and single-cell assays. *Nat. Protoc.* **1**, 2979–2987 (2006).
- 33. Thisse, C. & Thisse, B. High-resolution in situ hybridization to whole-mount zebrafish embryos. *Nat. Protoc.* **3**, 59–69 (2008).
- 34. Gosner, K. L. A simplified table for staging anuran embryos and larvae with notes on identification. *Herpetologica* **16**, 183–190 (1960).



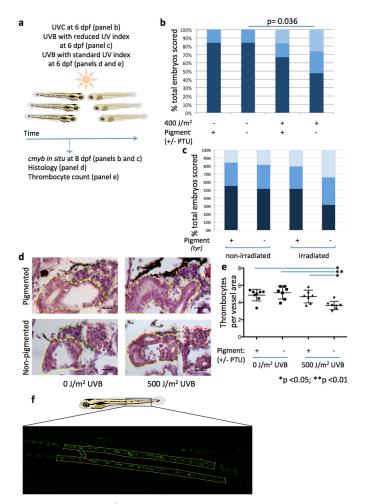
zebrafish kidney but dispensable for steady-state haematopoiesis. a, Top right panel, schematic of an embryo. The black boxed area contains the kidney and is enlarged in the other panels. Left panel containing 4 images, zebrafish positive for Tg(cdh17:GFP) (green, labelling the kidney tubule) and Tg(runx:mcherry) (red, labelling HSPCs) are depicted at 7 dpf. The boxed areas in the left images are enlarged in the corresponding fluorescence panels in the middle; these panels show the head kidney containing the haematopoietic marrow (indicated by the dashed outline). The white arrow highlights the melanocyte umbrella. Scale bars, 100 μm (bright field) and 50 µm (fluorescence). Bottom right panels, the kidney marrow of a 6-dpf larva positive for *Tg(mitfa:GFP)* (green, labelling melanocytes) and *Tg*(*runx:mcherry*) (red, labelling HSPCs) is shown from a lateral view (bottom right, large panel) and from an orthogonal view (transverse section; bottom right, small panel). The scale bar represents 20 μm . ${f b}$, Whole-mount in situ hybridization of cmyb at different time points, a representative larva at 5 dpf is shown. Arrows indicate (from cranial to caudal) the thymus, the kidney and the caudal haematopoietic tissue; the enlarged portion of the image (dashed box) shows the thymus and the kidney. The experiment was performed with n = 10 wild-type and 10 $mitfa^{-/-}$ larvae at 5 dpf and 10 wild-type and 7 $mitfa^{-/-}$ larvae at 7.5 dpf. c, Flow cytometric analysis of the percentage of HSPCs that were positive for *Tg(runx:mCherry)* as a proportion of live cells in the kidney marrow of adult fish. Data are mean \pm s.d.; n.s., not significant. d-f, Relative abundance of progenitors, myelomonocytes and lymphocytes in adult wild-type fish and mitfa, $roy^{-/-}$ and $casper^{-/-}$ pigment mutants as assessed by flow cytometry as previously described³⁰. **c-f**, n = 6 wild-type, 6 mitfa^{-/-}, 3 roy^{-/-} and 4 casper^{-/-} fish. Data are mean \pm s.d.; analysis by

ANOVA.





Extended Data Fig. 2 | Unprotected haematopoietic cells accumulate DNA damage after UV irradiation. a, Sorted cells that were positive for $\mathit{Tg(runx:mCherry)}$ after anti-CPD immunostaining. Scale bars, $10~\mu m$. b, Dot plot representation of data in Fig. 2b (quantification of immunostaining intensity per cell). Data are mean \pm s.d.; for statistics and P values please refer to Fig. 2b. c, Magnification of the thymus (dashed red outline) after anti-CPD immunostaining (counterstaining with haemalum) after UVB irradiation at 5 dpf in pigmented and non-pigmented larvae. Arrows from above indicate melanocytes; arrows from below indicate examples of nuclei with DNA damage.



Extended Data Fig. 3 | UV light is detrimental to exposed

haematopoiesis. a, Experimental layout. b, Reduction in HSPCs in larvae with and without PTU treatment as assessed by *cmyb* in situ hybridization after UVC irradiation. n = 26, 26, 24 and 19 larvae in from left to right. χ^2 test. c, Reduction in HSPCs in $tyr^{-/-}$ larvae and their pigmented siblings as assessed by *cmyb* in situ hybridization after UVB irradiation at a UV index of 20. n = 38, 33, 29 and 38 larvae from left to right. Results were not significant but note that the pigmented irradiated group seemed to retain more HSPCs in this experiment than in the preceding experiments with a higher UV index. d, Histology of the kidney marrow two days after irradiation. The yellow outline represents the kidney tubules, the red outline shows the aorta and the green outline indicates the haematopoietic marrow. Note the reduced area of the haematopoietic marrow in the non-pigmented, irradiated larvae (bottom right). Scale bars, 20 μm . **e**, Abundance of thrombocytes positive for Tg(CD41:GFP) two days after irradiation. Each data point represents the number of thrombocytes per vessel area in individual larvae. n = 8, 7, 7 and 7 larvae for treatment groups from left to right. Statistical significance was calculated using ANOVA with post hoc Bonferroni's multiple comparison test, data are mean \pm s.d. **f**, Schematic of the analysis of the number of thrombocytes that were positive for Tg(CD41:GFP). The boxed area represents the analysed area, in which the circulating cells were counted. The yellow outline represents the area of the vessel, which the number of circulating thrombocytes was normalized to.



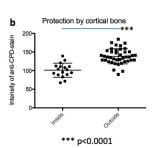
Extended Data Fig. 4 | **Sunlight exposure in fish living in the wild.** Example of small fish swimming in clear and shallow water on a sunny day (photo taken at Titisee, Germany, July 2016). Field of view approximately 25–30 cm wide.

Extended Data Fig. 5 | HSPCs relocated into the bone marrow before the transition to a terrestrial environment. a, Top, a developmental time line of D. tinctorius. Tadpoles were staged according to Gosner³⁴. Animals (from left to right) represent Gosner stage 25, 30, 41 and 42, as well as a

froglet five days after losing its tail. Stage-specific histology analyses are shown in panels \mathbf{b} - \mathbf{f} as indicated. Bottom, the habitat of D. tinctorius at the different developmental stages. \mathbf{b} - \mathbf{f} , Haematopoietic niches analysed by haematoxylin and eosin staining. Scale bars, $100~\mu m~(\mathbf{b}$ - \mathbf{f}) and $1~cm~(\mathbf{a})$.







Extended Data Fig. 6 | Cortical bone protects from UV-induced DNA damage. a, Paraffin section of a D. tinctorius hind leg (from Extended Data Fig., specimen e) after irradiation with UVB post mortem; the leg was severed from the body and irradiated with UVB. The dashed outline represents the cortical bone. Note the higher staining intensity of the anti-CPD antibody in nuclei within the muscle tissue compared to nuclei within the bone marrow. This part of the leg is not yet haematopoietic (compare to Extended Data Fig. 5e, which shows the haematopoietic marrow in the other leg), but contains chondrocytes. Note that even the chondrocyte nuclei closest to the cortical bone are stained much less than the cells outside the cortical bone (arrows from below and from above, respectively). The triangle represents the direction of the UV source; white tip towards UV source. Scale bar, 50 μ m. **b**, Quantification of grey scale values of nuclei inside (n = 17) and outside (n = 41) the cortical bone. Each data point represents the mean grey value of a 16 \times 16 pixel circle inside the nucleus, the difference is highly significant (unpaired Student's two-tailed *t*-test, P < 0.0001); data are mean \pm s.d.



Extended Data Fig. 7 | Experimental setup during irradiation. Fish were placed in a Petri dish inside the upper cardboard box to focus the light from above, because the Waldmann UV 801 BL unit has a curved

lamp carrier. The lower cardboard box was used to place the larvae at the recommended distance from the lamps.



Extended Data Table 1 \mid List of causes for lack of pigmentation

Unpigmented because of	Name	Mechanism
Chemical treatment	1-phenyl 2-thiourea (PTU)	Blocking of the enzyme tyrosinase ²⁹
		→ melanocytes do not contain melanin
Genetic cause	nacre (mitfa ^{w2})	Mutation in the transcription factor mitfa ³²
		→ melanocytes are absent
	sandy (tyr ^{tk20})	Mutation in the enzyme Tyrosinase ³³
		→ melanocytes do not contain melanin
	roy (mpv17 ^{a9})	Mutation in the mpv17 mitochondrial inner
		membrane protein ³⁴
		→ iridophores are absent
	casper (mitfa $^{w2/w2}$; mpv17 $^{a9/a9}$)	double homozygous for nacre and roy34
		→ melanocytes and iridophores are
		absent



GLI1-expressing mesenchymal cells form the essential Wnt-secreting niche for colon stem cells

Bahar Degirmenci^{1,4}, Tomas Valenta^{1,2,4*}, Slavica Dimitrieva³, George Hausmann¹ & Konrad Basler^{1*}

Wnt-β-catenin signalling plays a pivotal role in the homeostasis of the intestinal epithelium by promoting stem cell renewal^{1,2}. In the small intestine, epithelial Paneth cells secrete Wnt ligands and thus adopt the function of the stem cell niche to maintain epithelial homeostasis^{3,4}. It is unclear which cells comprise the stem cell niche in the colon. Here we show that subepithelial mesenchymal GLI1expressing cells form this essential niche. Blocking Wnt secretion from GLI1-expressing cells prevents colonic stem cell renewal in mice: the stem cells are lost and, as a consequence, the integrity of the colonic epithelium is corrupted, leading to death. GLI1-expressing cells also play an important role in the maintenance of the small intestine, where they serve as a reserve Wnt source that becomes critical when Wnt secretion from epithelial cells is prevented. Our data suggest a mechanism by which the stem cell niche is adjusted to meet the needs of the intestine via adaptive changes in the number of mesenchymal GLI1-expressing cells.

Intestinal epithelial stem cells (IES cells) serve as a paradigm for adult stem cells. Stem-cell populations reside in niches—microenvironments that regulate the participation of stem cells in tissue generation, maintenance and repair^{5,6}. Wnt ligands are a critical component provided by intestinal stem cell niches.

In the small intestine epithelium, the primary niche comprises terminally differentiated Paneth cells, which act as a source of Wnt signalling molecules^{3,7,8}. In vitro studies have shown that Wnts provided by Paneth cells are sufficient to maintain epithelial renewal. It was therefore unexpected that homeostasis of the small intestine epithelium tolerates the loss of Paneth cells or the absence of epithelial Wnt secretion^{1,9–11}. In contrast to the situation in the small intestine, the identity of the cells that constitute the Wnt-producing niche in the colon has remained a mystery. Unlike the small intestine, colonic crypts do not contain Paneth cells or any other Wnt-secreting epithelial cell type^{7,8,12}. Analysis of mice in which Wnt secretion was globally blocked confirmed that there is an extra-epithelial Wnt-producing niche in the small intestine¹. Although this work suggested that cells in the subepithelial mesenchyme comprise a niche, it raised crucial questions regarding the identity of this population and how the niche is regulated. This issue is of substantial importance for intestinal stem cell research.

GLI1⁺ cells express Wnts and are located near the bases of intestinal crypts¹ (Extended Data Fig. 1a–c). Given the accumulating evidence that Wnt proteins in the intestinal stem-cell niche act as a short range signal, we focused on this cell population¹³. To identify the mesenchymal origin of the GLI1⁺ cells, we performed a lineage-tracing experiment using *Gli1-Cre^{ERT2};Pdgfra^{EGFP};lox-STOP-lox-tdTomato* mice, in which all GLI1⁺ cells are marked with tdTomato. Consistent with a mesenchymal origin, all GLI1⁺ cells were also positive for PDGFRA (Extended Data Fig. 1c, d), which is a marker of undifferentiated mesenchymal cells¹⁴. We further confirmed by lineage tracing that GLI1⁺ cells are long-lived (Extended Data Fig. 2a). Like GLI1⁺ mesenchymal cells from numerous other tissues, intestinal GLI1⁺ mesenchymal cells can be differentiated in vitro into smooth muscle, osteoblasts and adipocytes^{15,16} (Extended Data Fig. 2b, c).

To gain insight into the necessity of Wnt secretion from GLI1⁺ cells for intestinal homeostasis and stem cell maintenance, we combined a conditional *Wntless (Wls)* allele (*Wls^{flox}*) with an inducible GLI1⁺ cell-specific Cre driver (*Gli1-Cre^{ERT2}*), referred to henceforth as *Gli1-Wls^{cRO}* (full genotype: *Gli1-Cre^{ERT2}*; *Wls^{flox/flox}*). WLS is a key protein required for the secretion of all Wnt ligands, and eliminating WLS blocks Wnt secretion^{17,18}.

If *Wls* is eliminated in the colonic epithelia, the crypts remain intact; in striking contrast, the majority of the colonic crypts in *Gli1-WlscKO* mice were lost 14 days after Cre induction (Fig. 1a). Twenty-one days after Cre induction, the colonic epithelium had been completely destroyed (Fig. 1a). The stem cell marker LGR5 had disappeared at a stage during which colon morphology was still normal (Fig. 1b, Extended Data Fig. 3a). Wnt pathway activity in the colonic epithelium of *Gli1-WlscKO* mice was markedly reduced, as shown by a decrease in the expression of Wnt targets such as *Axin2*, *CyclinD1* (*Ccnd1*), *Cd44*, *Ephb2* and the stem cell markers *Tnfrsf19* (*Troy*)¹⁹ and *Ascl2* (Fig. 1c, Extended Data Fig. 4b, c). These observations demonstrate that GLI1⁺ cells are the indispensable, long-sought Wnt source that comprises the niche of colon IES cells.

The dependency of colon IES cells solely on Wnts secreted by GLI1⁺ cells was intriguing, considering that the small intestinal crypts of Gli1-Wls^{cKO} mice were similar to those of control mice (Fig. 1a). If GLI1⁺ cells represent the putative second niche, beside Paneth cells, in the small intestine, small intestinal morphology should be disrupted when Wnt secretion is blocked both in the epithelium and in GLI1⁺ cells. Indeed, the combination of *Gli1-Cre^{ERT2}* and the intestinal epithelial cell-specific *Villin-Cre^{ERT2}* drivers, together with the conditional *Wls* allele (henceforth referred to as $Villin+Gli1-Wls^{cKO}$ (genotype: $Gli1-Cre^{ERT2}$; $Villin-Cre^{ERT2}$; $Wls^{flox/flox}$) resulted in mice in which the small intestinal epithelium was totally destroyed. Sixteen days after induction of recombination, proliferation-deficient crypts were seen in the duodenum of Villin+Gli1-WlscKO mice (Fig. 1d). Twenty-one days after Cre induction, no crypts were present (Fig. 1d). Crypts in the colons of *Villin+Gli1-WlsckO* mice were indistinguishable from those of Gli1-Wls^{cKO} mice (Extended Data Fig. 3a). Preempting the subsequent deterioration of the tissue, the stem cell markers²⁰ OLFM4 and LGR5 disappeared from the bases of crypts in the duodenum of Villin+Gli1- Wls^{cKO} mice after 14 days, when the morphology of the epithelium still appeared normal (Fig. 1e, Extended Data Figs. 3b, 4a). Notably, the addition of exogenous Wnts (WNT3A and WNT2B) delayed the onset of epithelial phenotypes in the duodenum and colon of Villin+Gli1-WlsckO mice (Extended Data Fig. 4d, e). Together, these data strongly indicate that subepithelial GLI1⁺ cells can compensate for the absence of Wnt secretion from Paneth cells and thus that GLI1⁺ cells may represent the reserve niche for IES cells in the small intestine.

If mesenchymal GLI1-expressing cells constitute a Wnt-producing niche, then these cells should be able to maintain the viability of colonic organoids in vitro^{21,22}. To test this prediction, we isolated GLI1⁺ cells from the duodenum and colon of *Gli1-Cre^{ERT2}*, *lox-STOP-lox-tdTo-mato* mice 24 h after Cre induction by sorting for tdTomato⁺ cells.

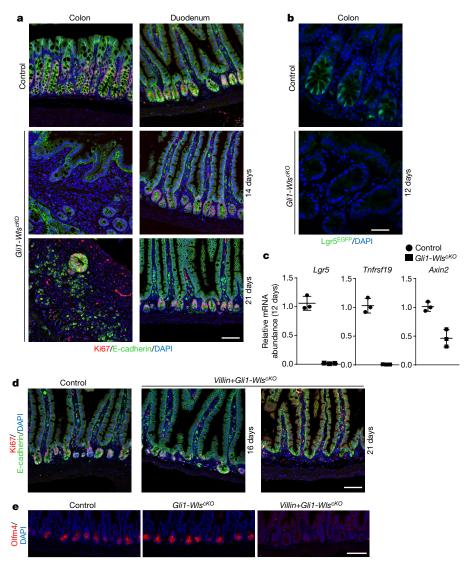


Fig. 1 | GLI1⁺ cells constitute an essential Wnt niche that supports IES cells in the colon and serve as a reserve niche in the small intestine.

a, Blocking Wnt secretion from GLI1⁺ cells leads to loss of proliferation activity (marked by Ki67) and loss of crypts from the proximal colon.

b, Disappearance of IES cells marked by LGR5 precedes reduced cell proliferation in the colon. c, Expression of *Axin2* decreases and the expression of IES cell markers (*Lgr5*, *Tnfrsf19*) is strongly reduced

in the colon when Wnt secretion is ablated in GLI1⁺ cells (real-time quantitative PCR with reverse transcription (RT–qPCR), n=3 biologically independent animals, mean \pm s.d.). **d**, Only simultaneous blocking of Wnt secretion from both the epithelium and GLI1⁺ cells results in loss of crypts in the small intestine. **e**, Expression of the stem cell marker OLFM4 is lost in small intestinal crypts of *Villin+Gli1-Wls^{cKO}* mice, but is unchanged in *Gli1-Wls^{cKO}* mice. Scale bars, 50 μ m.

Colonic organoids require the addition of exogenous Wnts to the culture medium to be able to self-renew and proliferate²². Colonic organoids cocultured with GLI1⁺ cells could grow normally without exogenous WNT3a and were morphologically similar to colonic organoids cultured in medium supplemented with exogenous Wnts (Extended Data Fig. 5b–d). Consistent with secretion of active Wnts by GLI1⁺ cells, duodenal organoids shifted their morphology from protrusions to spheroids upon coculturing with GLI1⁺ cells (Extended Data Fig. 5a). The notion that a critical role of GLI1⁺ cells is to secrete Wnts was further confirmed by coculturing GLI1⁺ (tdTomato⁺) cells with duodenal *Villin-Wls^{cKO}* organoids. These organoids die without addition of external Wnts¹. Coculture with GLI1⁺ cells restored the growth of *Villin-Wls^{cKO}* duodenal organoids (Extended Data Fig. 5c, e), and the GLI1⁺ cells were in close proximity to the epithelial organoids (Extended Data Fig. 5c).

To further characterize the GLI1⁺ cell population, we used single-cell RNA sequencing (scRNA-seq) on sorted colonic GLI1⁺ (tdTomato⁺) cells. Analysis of scRNA-seq data revealed eight distinct clusters (C1–C8; Fig. 2a, Extended Data Fig. 6a, b). Cells positive for the mesenchymal markers *Acta2* and *Vim* are present in all clusters; abundant

Acta2 expression overlapping with Myh11 expression characterizes C2 as a putative myofibroblast cluster (Extended Data Figs. 6c, 7a). Pdgfra transcripts were present in all clusters, consistent with the overlapping expression of GLI1 determined in vivo (Extended Data Figs. 6d, 7a). Another broad mesenchymal marker, Cd34, was enriched mainly in clusters C1 and C3 (Extended Data Figs. 6e, 7a). Indeed, ACTA2 and CD34 are expressed in a much larger part of the colonic mesenchyme than GLI1²³ (Extended Data Fig. 1a, b). Foxl1, which is expressed by a fraction of subepithelial mesenchymal cells²⁴, was found in clusters C2 and C4, where it overlaps with Acta2 and Myh11 (Extended Data Fig. 7a, b). Wnt2, Wnt2b and Wnt4 transcripts were enriched in clusters C1 and C3. Rspo3, which encodes a potent stimulator of Wnt signalling, occupied cluster C1 (Extended Data Figs. 6f, 7a, c).

Overall, the comparative expression analysis revealed that GLI1⁺ cells constitute a heterogeneous population. Profiling of the highly expressed genes of each cluster indicated that expression of *Rbp1* and *Sfrp1* roughly marks the Wnt-expressing clusters C1 and C3, respectively (Fig. 2b). Using the RBP1 and SFRP1 markers in combination with WNT4, we could assess the location of these clusters in the colonic tissue. Immunohistochemical analysis revealed not only a

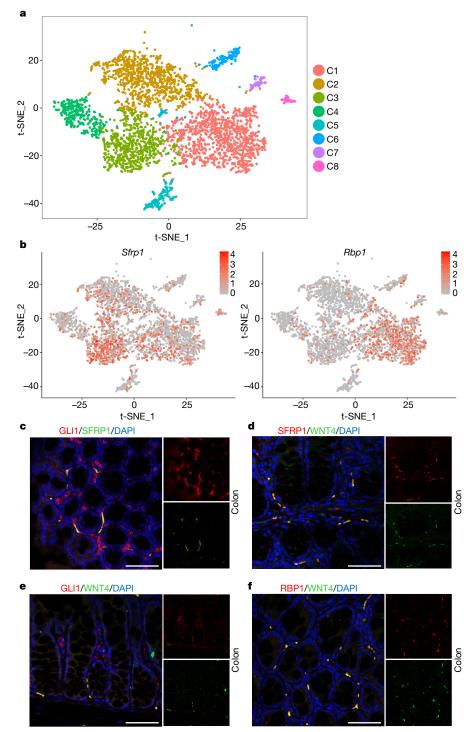


Fig. 2 | GLI1 $^+$ cells constitute a heterogenous population with certain subsets localized in close proximity to epithelial crypts. a, GLI1 $^+$ cells constitute eight distinct clusters (C1 $^-$ C8) as revealed by unbiased t^- distributed stochastic neighbour embedding (t-SNE) clustering analysis of scRNA-seq data from sorted colonic GLI1 $^+$ cells. Each dot represents an

individual cell. **b**, Expression of *Sfrp1* and *Rbp1* in distinct populations of GLI1 $^+$ cells. **c**–**f**, A subpopulation of GLI1 $^+$ cells co-expressing SFRP1 and WNT4 is localized in close proximity to intestinal colon crypts. WNT4 $^+$ pericryptic cells are marked by RBP1. Scale bars, 50 μ m.

close proximity of WNT4⁺ cells to the base of the colonic crypts, but also a spatial overlap of those cells with RBP1 and SFRP1 expression (Fig. 2c-f). While our analysis cannot exclude the existence of GLI1-negative cells that also secrete Wnt ligands, we know from our genetic experiments that such cells are not able to compensate for a lack of Wnt secretion from GLI1⁺ cells.

In an unperturbed state, GLI1⁺ cells serve as the essential niche for colon IES cells and as a reserve niche for small intestine IES cells. To analyse how this mechanism operates in a challenged condition,

we induced colitis, an inflammatory situation in which extra regeneration is required. Mice were treated with 2.5% DSS (dextran sulfate sodium) for 5 days and then allowed to recover for 1, 3 or 5 days before being killed for analysis of GLI1⁺ cells (Extended Data Fig. 8a). We observed a significant increase in the number of GLI1⁺ cells in mice treated with DSS when compared with control mice ($P_{3 \text{ days}} = 3 \times 10^{-5}$; $P_{5 \text{ days}} = 3 \times 10^{-6}$) (Fig. 3a, b, Extended Data Fig. 8b). GLI1⁺ cells thus not only act as an essential niche for the stem cells, but may also contribute to the restoration of colon homeostasis during recovery from

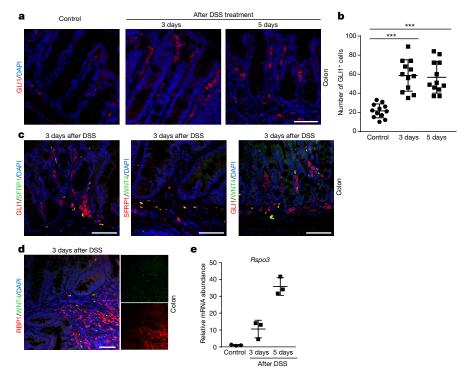


Fig. 3 | A specific sub-population of GLI1⁺ cells is enriched during regeneration following DSS-mediated epithelial damage. a, Increase in GLI1⁺ cells during recovery from DSS-mediated colonic damage. b, Quantification of GLI1⁺ cells in the colon of control mice (n=4 independent mice), and in mice recovering from DSS treatment (n=4 for each time point); mean \pm s.d. *** $P \le 0.001$ (t-test, one-sided);

 $P_{3 \text{ days}} = 3 \times 10^{-5}, P_{5 \text{ days}} = 3 \times 10^{-6}.$ c, The GLI1+ cells that are enriched upon DSS-induced damage represent a subpopulation distinct from the SFRP1+WNT4+ cells. d, The RBP1+ subpopulation is enriched upon DSS-induced damage. e, RSpo3 expression increased during regeneration of the colon after DSS-induced damage (RT–qPCR, n=3 biologically independent mice, mean \pm s.d.). Scale bars, 50 μm .

DSS treatment by governing the size of the niche. The extra GLI1⁺ cells observed do not correspond to myeloid cells, as assessed by immunohistochemical analysis of the myeloid markers F4/80 and CD11C (Extended Data Fig. 8c). To determine which of the GLI1⁺ subpopulations had expanded, we took advantage of our scRNA-seq data analysis. While the number of SFRP1⁺ cells did not noticeably increase, RBP1⁺ cells were more abundant, consistent with the increase in GLI1⁺ cells (Figs. 2c–f, 3c, d). Although the number of WNT4⁺ cells did not show a marked change, the expression of *Rspo3* increased notably. RSPO3⁺ cells occupy the same cluster (C1) as RBP1⁺ cells (Fig. 3e, Extended Data Fig. 7a). Together, these data identify RBP1⁺ cells as the subset of GLI1⁺ cells that is enlarged during the recovery period after DSS treatment.

In the small intestine, Villin-WlscKO mice (in which epithelial Wnt production is blocked) showed an increase in the number of GLI1⁺ cells compared with control mice (Extended Data Fig. 9a, b). Thus, subepithelial GLI1⁺ cells respond to epithelial Wnt signal loss by increasing the niche size to compensate for the role of epithelial (Paneth) cells. We determined that GLI1 expression is responsive to Hedgehog pathway stimulation by treating the cells with recombinant SHH ligand or the compound Smoothened agonist (SAG) (Extended Data Fig. 9c, d). Notably, expression of epithelial Shh was moderately increased when epithelial Wnt secretion was blocked, whereas expression of Ihh was unchanged (Extended Data Fig. 9e, f). These findings suggest a convenient mechanism to compensate for the loss of epithelial Wnt secretion in the duodenum: when homeostasis of the small intestinal epithelium is disrupted, increased Hedgehog signalling in the epithelium may serve as a regulatory cue to augment the number of GLI1⁺ cells to meet the increased Wnt demands of stem cells.

Our work has revealed the identity of the source of Wnt ligands in the colon, which is required for colon homeostasis. We have demonstrated that $\mathrm{GLI1}^+$ cells constitute a critical Wnt-producing stem cell niche that promotes the self-renewal of colonic IES cells. We also found that the number of $\mathrm{GLI1}^+$ cells, and hence the size of the niche, increases during tissue regeneration.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0190-3.

Received: 10 September 2017; Accepted: 4 May 2018; Published online: 06 June 2018

- Valenta, T. et al. Wnt ligands secreted by subepithelial mesenchymal cells are essential for the survival of intestinal stem cells and gut homeostasis. Cell Rep. 15, 911–918 (2016).
- Fevr, T., Robine, S., Louvard, D. & Huelsken, J. Wnt/β-catenin is essential for intestinal homeostasis and maintenance of intestinal stem cells. *Mol. Cell. Biol.* 27, 7551–7559 (2007).
- Sato, T. et al. Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. Nature 469, 415–418 (2011).
- Snippert, H. J. et al. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. Cell 143, 134–144 (2010).
- Barker, N. et al. Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature 449, 1003–1007 (2007).
- Scadden, D. T. The stem-cell niche as an entity of action. Nature 441, 1075–1079 (2006).
- Barker, N. Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration. Nat. Rev. Mol. Cell Biol. 15, 19–33 (2014).
- Farin, H. F., Van Es, J. H. & Clevers, H. Redundant sources of Wnt regulate intestinal stem cells and promote formation of Paneth cells. Gastroenterology 143, 1518–1529.e7 (2012).
- Durand, A. et al. Functional intestinal stem cells after Paneth cell ablation induced by the loss of transcription factor Math1 (Atoh1). Proc. Natl Acad. Sci. USA 109, 8965–8970 (2012).
- Kim, T. H., Escudero, S. & Shivdasani, R. A. Intact function of Lgr5 receptorexpressing intestinal stem cells in the absence of Paneth cells. *Proc. Natl Acad.* Sci. USA 109, 3932–3937 (2012).
- Kabiri, Z. et al. Stroma provides an intestinal stem cell niche in the absence of epithelial Wnts. *Development* 141, 2206–2215 (2014).
- Sasaki, N. et al. Reg4⁺ deep crypt secretory cells function as epithelial niche for Lgr5⁺ stem cells in colon. *Proc. Natl Acad. Sci. USA* 113, E5399–E5407 (2016).
- 13. Farin, H. F. et al. Visualization of a short-range Wnt gradient in the intestinal stem-cell niche. *Nature* **530**, 340–343 (2016).



- Morikawa, S. et al. Prospective identification, isolation, and systemic transplantation of multipotent mesenchymal stem cells in murine bone marrow. J. Exp. Med. 206, 2483–2496 (2009).
- Kramann, R. et al. Perivascular Gli1⁺ progenitors are key contributors to injury-induced organ fibrosis. Cell Stem Cell 16, 51–66 (2015).
- Kramann, R. et al. Adventitial MSC-like cells are progenitors of vascular smooth muscle cells and drive vascular calcification in chronic kidney disease. Cell Stem Cell 19, 628–642 (2016).
- Bänziger, C. et al. Wntless, a conserved membrane protein dedicated to the secretion of Wnt proteins from signaling cells. Cell 125, 509–522 (2006).
- Degirmenci, B., Hausmann, G., Valenta, T. & Basler, K. Wnt ligands as a part of the stem cell niche in the intestine and the liver. *Prog. Mol. Biol. Transl. Sci.* 153, 1–19 (2018).
- Fafilek, B. et al. Troy, a tumor necrosis factor receptor family member, interacts with Lgr5 to inhibit Wnt signaling in intestinal stem cells. Gastroenterology 144, 381–391 (2013).
- van der Flier, L. G., Haegebarth, A., Stange, D. E., van de Wetering, M. & Clevers, H. OLFM4 is a robust marker for stem cells in human intestine and marks a subset of colorectal cancer cells. Gastroenterology 137, 15–17 (2009).
- Sato, T. et al. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature 459, 262–265 (2009).
- Sato, T. et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. Gastroenterology 141, 1762–1772 (2011).
- Stzepourginski, I. et al. CD34⁺ mesenchymal cells are a major component of the intestinal stem cells niche at homeostasis and after injury. Proc. Natl Acad. Sci. USA 114, E506–E513 (2017).
- 24. Aoki, R. et al. Foxl1-expressing mesenchymal cells constitute the intestinal stem cell niche. *Cell. Mol. Gastroenterol. Hepatol.* **2**, 175–188 (2016).

Acknowledgements We thank O. Sansom for providing us with the *Gli1-Cre^{ERT2}* and *Pdgfra^{EGFP}* strains and for suggestions; S. Robine for the *Villin-Cre^{ERT2}* strain; F. Greten, L. Sommer, M. Aguet and G. Christofori for comments; members of

the Basler laboratory, in particular C. Cantù, V. S. Salazar and D. Zimmerli, for discussions; E. Escher, E. Tuncer, L. Zurkirchen and V. Parfejevs for technical help; J. Duarte, C. Ewald and A. Henning for help with cell sorting; and C. Aquino and the Functional Genomics Center Zurich for performing scRNA-seq. This work was supported by the Swiss National Science Foundation, the Swiss Cancer League, the University of Zurich Research Priority Program (URPP) 'Translational Cancer Research' and the Kanton of Zürich. T.V. is supported by Czech Science Foundation grant 18-21466S and is a fellow of the URPP Translational Cancer Research.

Reviewer information Nature thanks C. Kuo, L. Samuelson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.D. and T.V. designed and performed experiments and collected and analysed data. T.V. initiated and conceived the research. B.D. and T.V. wrote the manuscript and performed all experiments together with data analysis. S.D. analysed scRNA-seq data. G.H. discussed the data and assisted with manuscript preparation. K.B. initiated and supported the research, discussed the data and assisted with manuscript preparation.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0190-3.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0190-3.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to T.V. or K.B. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mouse experiments. Mouse experiments were performed in accordance with Swiss guidelines and approved by the Veterinarian Office of Kanton Zürich, Switzerland. To achieve conditional deletion of *Wls* (*Wls*^{cKO}) the *Wls* conditional allele¹ was combined with either the Gli1-Cre^{ERT2} or Villin-Cre^{ERT2} driver^{25,26}. For the determination of LGR5 expression in vivo, the knock-in allele Lgr5-EGFP-IRES-Cre^{ERT2} was used⁵. To induce Cre-mediated recombination, tamoxifen (Sigma) was injected (80 mg/kg) intraperitoneally for 5 consecutive days. The day of the first injection was counted as day 0. External active mouse Wnts were applied as described with the following modification: a 1:1 mixture of mWnt3a (Abcam, R&D) and mWnt2b (R&D) was used, and 200 $\mu g/kg$ of the Wnt mixture was applied per injection. To trace GLI1⁺ cells, the *Gli1-Cre^{ERT2}* driver was combined with a *lox-STOP-lox* $tdTomato^{27}$ transgene (ROSA26Sor^(CAG-tdTomato)). To mark PDGFRA expression in vivo, the knock-in allele *Pdgfra*^{EGFP} was used²⁸. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment. Both males and females at age 8-12 weeks were used.

RNA isolation, cDNA synthesis, quantitative real-time PCR. Freshly isolated intestine was lysed in Tri-Reagent (Sigma) and total RNA was isolated per the manufacturer's instructions. In the experiments in which intestinal epithelial cells were used, they were isolated as described¹. For cDNA synthesis, 1 µg total RNA was used and cDNA was synthesized using the RNA to cDNA EcoDry synthesis kit (TaKaRa/Thermofisher). Quantitative real-time PCR reactions were performed minimally in three independent biological replicates, each with technical triplicates using the Applied Biosystems SYBR Green Kit and monitored by the QuantStudio3 system (Applied Biosystems). Representative results are shown in the Figures. Primer sequences for Acth, Gapdh, Sdha, Axin2, Ascl2, Lgr5 and Tnfrsf19 were as previously published¹. For other genes, the sequences are:

Bmp4: GAGGAGTTTCCATCACGAAGA, GCTCTGCCGAGGAGATCA; Cd44: TCCTTCTTTATCCGGAGCAC, ACGTCTCCTGCTGGGTAGC; Ccnd1: TTTCTTTCCAGAGTCATCAAGTGT, TGACTCCAGAAGGGCTTCAA; Ephb2: CCTGATGAACCTTCACAACAAC, TCTTGTTTCAAGAAGCGCTTTAC; Gli1: CCAAGCCAACTTTATGTC, AGCCCGCTTCTTTGTTAA; Ihh: GGCTTCG ACTGGGTGTATTA, CGGTCCAGGAAAATAAGCAC; Ptch1: AAAGAAC TGCGGCAAGTT, CTTCTCCTATCTTCTGAC; Ptch2: CCCGTGGTA ATCCTCGTGGCC, TCCATCAGTCACAGGGGCAAA; Rspo1: CGACATGA ACAAATGCATCA, CTCCTGACACTTGGTGCAGA; RSpo2: GCCCATCAGG GTATTATGGA, TCACAGTTTTCTATTCTGCATCG; RSpo3: TCAAAGGGA GACCGAGGA, TGCTGTCAGAGGAGGAGCTT; Shh: CCAATTACAA CCCCGACATC, GCATTTACTTTTTTCTCATCAC; Wnt2: CCTGATGAACC TTCACAACAAC, CTTTGTTTCAAGAAGCGCTTTAC; Wnt2b: GGGCCC TCATGAACTTACAC, CCACTCACACCGTGACACTT; Wnt4: ACTGGAC TCCCTCCCTGTCT, TGCCCTTGTCACTGCAAA.

For RT–qPCR, samples were measured in triplicate and average cycle threshold values were quantified relative to Gapdh or three reference genes (Actb, Gapdh and Sdha) using the $\Delta\Delta$ CT method¹.

Histology, immunohistochemistry and immunocytochemistry. Freshly isolated tissue was fixed in 4% paraformaldehyde, dehydrated and mounted in paraffin using standard protocols. Material for frozen sections was processed and embedded in OCT (Tissue-Tek; Sakura Finetek) according to standard protocols. Cells were cultured in 8-well staining chambers (Laboratory-Tek Chamber Slide System-Permanox, Nunc) and were fixed in 4% paraformaldehyde. Oil Red O and Alizarin Red S staining was carried out according to standard protocols. Standard immunohistochemical protocols were performed with the following antibodies: mouse anti-Acta2 (Sigma), mouse anti-E-cadherin (BD Transduction Laboratories), rabbit anti-Ki67 (Abcam), rabbit anti-Gli1 (Novus), rabbit anti-Shh (Abcam), rabbit anti-Olfm4 (Cell Signaling), rabbit anti-RFP (Rockland Immunochemicals), rabbit anti-Sfrp1 (Abcam), goat anti-Sfrp1 (Abcam), rabbit anti-Rbp1 (Abcam), goat anti-Wnt4 (R&D Systems), Armenian hamster anti-Cd11c (Biolegend) and rat anti-F4/80 (Bio-rad). Secondary antibodies were anti-rabbit, anti-mouse, antigoat, anti-rat and anti-Armenian hamster antibodies conjugated with Alexa (A488, A594 or A647) from ThermoFisher Scientific or Abcam. For Shh detection, the Vectastain ABC kit and DAB peroxidase substrate kit (both Vector) were used. Images were taken using a Leica LSM 710 or Leica SP8 confocal microscope, and processed using ImageJ (FIJI) and AdobePhotoshopCS6 software.

Intestinal organoids, mesenchymal cell culture and cell sorting. Intestinal organoids were generated and cultured as described described from crypts of $Villin-Wls^{cKO}$ or control mice injected five times with tamoxifen removed 7 days after the first tamoxifen application. Organoids were cultivated in medium supplemented as follows: ENR = EGF (Gibco/Thermofisher) 50 ng/ml, Noggin (Sigma) 100 ng/ml, R-spondin1 (Sigma) 500 ng/ml; ENRW = ENR + 30% Wnt3a conditioned medium.

Intestinal mesenchymal cells were isolated as described³⁰; Gentle Cell Dissociation Reagent (StemCell Technologies) was used for removal of

epithelial cells. Intestinal mesenchymal cells were cultivated in MesenCult (StemCell Technologies) with MSC Stimulatory supplements and for the first 5 days also with MesenPuro (StemCell Technologies). Cre-mediated recombination in vitro was induced by addition of 4-(Z)-hydroxytamoxifen (Sigma) (500 ng/ml) for 12 h. Smoothened Agonist (SAG; Sigma) or recombinant highactive human SHH (R&D) was used to activate the Hedgehog pathway after 12 h of serum starvation in complete medium for 20 h. Differentiation was induced by MesenCult Adipogenic Stimulatory kit (StemCell Technologies) or MesenCult Osteogenic Stimulatory kit (StemCell Technologies) for 21 days as described by the manufacturer. Differentiation towards smooth muscle cells was achieved using cultivation in alphaMEM (Glutamax) (Gibco/Thermofisher), supplemented with 20% FBS, penicillin/streptomycin, 10 ng/ml TGFβ (StemCell Technologies) and 5 ng/ml PDGF-bb (StemCell Technologies). For cell sorting, mice were injected with tamoxifen 24 h before cell isolation; intestinal mesenchymal cells were isolated as described above and sorted for tdTomato⁺ from Gli1-Cre^{ERT2}, lox-STOP-loxtdTomato mice. Sorting was performed at the Cytometry/Flow Cytometry facility of the University of Zurich using a FACSAria III cell sorter (BD Biosciences) (Extended Data Fig. 5f, g). Thirty thousand sorted cells were mixed with 100 crypts and plated directly in Matrigel.

scRNA-seq. The Chromium Single Cell 3' v2 Reagent Kit (10x Genomics) was used in the succeeding steps. Five thousand cells were loaded on a Chromium Single Cell Chip to partition the cells into nanolitre-scale Gel Bead In-EMulsions (GEMs). The cells dissolved upon contact with the Single Cell 3' Gel Bead in a GEM and primers containing (i) an Illumina R1 sequence (read 1 sequencing primer), (ii) a 16-nt10x Barcode, (iii) a 10-nt Unique Molecular Identifier (UMI) and (iv) a poly-dT primer sequence were released and mixed with cell lysate and Master Mix. The GEMs were reversed transcribed and barcoded, and full-length cDNA was produced from poly-adenylated mRNA. After incubation, the GEMs were broken and the pooled fractions recovered.

Silane magnetic beads were used to remove leftover biochemical reagents and primers from the post-GEM reaction mixture. Full-length, barcoded cDNA was then amplified by PCR to generate sufficient amounts for library construction.

Libraries were generated by ligating P5 and P7 Illumina adapters followed by size selection. The quality and quantity of the enriched libraries were validated using Tapestation (Agilent).

Sequencing was performed on the Illumina HiSeq 2500 v4 system. Reads were quality-checked with FastQC. (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

De-multiplexing, collapsing of unique molecular identifiers (UMIs) and alignment of reads to the *Mus musculus* transcriptome (GRCm38.p5 Ensembl release 89) were performed using the Cellranger toolkit (version 2.0.2) provided by 10x Genomics. To exclude low quality cells, we filtered out cells for which fewer than 500 genes were detected, leaving 4,464 cells for further analyses. All genes that were not detected in at least five cells were also discarded, leaving 15,190 genes.

The further downstream analysis was performed with Seurat R package (2.1.0) (http://satijalab.org/seurat/). Library-size normalization was done on the UMI-collapsed gene expression values for each cell barcode by scaling by the total number of transcripts and multiplying by 10,000. The data were then natural-log transformed for all downstream analyses. The detection of highly variable genes was based on the average expression and dispersion for each gene such that genes were placed into bins, and then z-scores for dispersion were calculated within each bin. This procedure identified 2,008 genes as variable across the dataset.

The dimensionality of the dataset was reduced using principal component analysis (PCA). To identify statistically significant principal components (PCs) we first used a modified randomization approach. In brief, we performed 500 PCAs on the input data, randomly permuting a subset of data (1% of the genes) in each analysis to estimate a null distribution of scores for every gene. For the PCs that exhibited strong enrichment of low *P* value genes, we examined the standard deviations using the Seurat function PCElbowPlot. Based on the elbow in this graph, we selected 12 PCs for downstream analyses. Unbiased clustering of the cells was performed using a shared nearest neighbour (SNN) modularity optimization-based clustering algorithm, as implemented in the FindClusters function in Seurat. Setting a resolution of 0.3 identified nine cell populations. Differentially expressed genes (markers) for each of the cell populations were identified using Wilcoxon rank sum test.

Statistics and reproducibility. The experiments in Fig. 1a, d and Extended Data Fig. 1a, b were repeated three times with four mice per time point. The experiments in Fig. 1b, e and Extended Data Fig. 4a were repeated twice with two mice. The experiments in Fig. 2c—f were repeated three times with two mice. The experiments in Fig. 3a, c, d and Extended Data Fig. 8b, c were repeated twice with four mice per time point. The experiments in Extended Data Fig. 1a, b were repeated twice with two mice. The experiment in Extended Data Fig. 1d was repeated in two independent experiments (three replicates per experiment). The experiment in Extended Data Fig. 2a was done twice with two mice. The experiments in Extended Data Fig. 2b, c were performed in three independent experiments (three replicates per



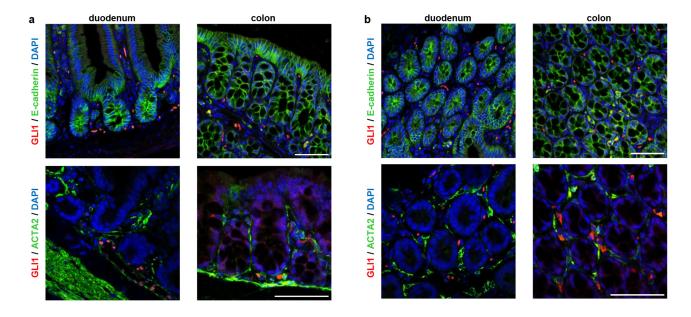
experiment). The experiments in Extended Data Fig. 4d, e were performed twice with two mice for each condition. The experiments in Extended Data Fig. 5a, c were repeated in two independent experiments (three replicates per experiment). The experiment in Extended Data Fig. 9a was repeated three times with three mice. The experiment in Extended Data Fig. 1d was repeated in two independent experiments (three replicates per experiment). The experiment in Extended Data Fig. 9e was repeated five times with two mice.

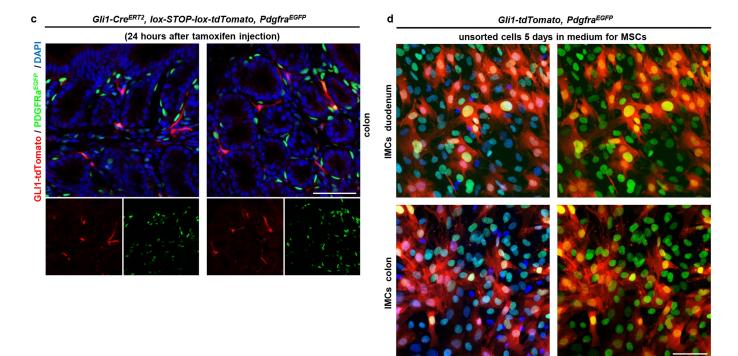
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The single-cell RNA-seq dataset generated during the current study is available in the GEO (Gene Expression Omnibus) database repository under accession number GSE113043. Source data are provided for the graphs in the following panels: Figs. 1c, 3b, e, Extended Data Figs. 4b, c, 5d, e, 9b, c, f. Extended versions of the legends of Figs. 1–3 are available as Supplementary Information.

The other datasets generated and/or analysed during the current study are available from the corresponding authors upon reasonable request.

- Ahn, S. & Joyner, A. L. Dynamic changes in the response of cells to positive hedgehog signaling during mouse limb patterning. Cell 118, 505–516 (2004).
- el Marjou, F. et al. Tissue-specific and inducible Cre-mediated recombination in the gut epithelium. Genesis 39, 186–193 (2004).
- 27. Madisen, L. et al. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci.* **13**, 133–140 (2010).
- Hamilton, T. G., Klinghoffer, R. A., Corrin, P. D. & Soriano, P. Evolutionary divergence of platelet-derived growth factor alpha receptor signaling mechanisms. Mol. Cell. Biol. 23, 4013–4025 (2003).
- Sato, T. & Clevers, H. Primary mouse small intestinal epithelial cell cultures. Methods Mol. Biol. 945, 319–328 (2013).
- Koliaraki, V. & Kollias, G. Isolation of intestinal mesenchymal cells from adult mice. *Bio Protoc.* 6, e1940 (2016).

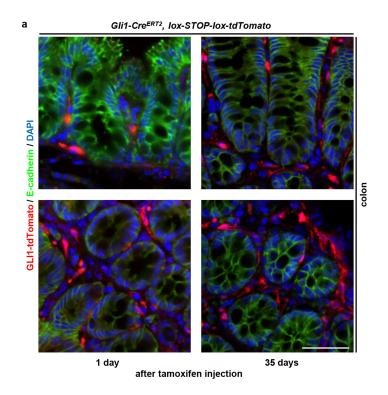


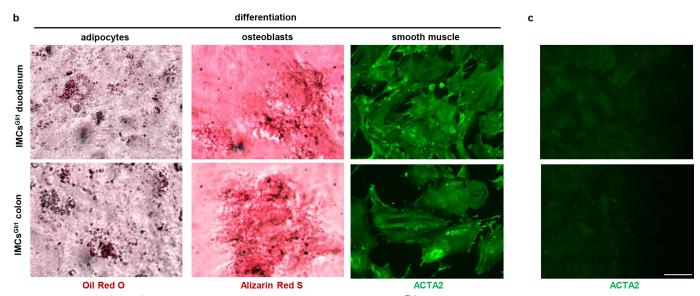


Extended Data Fig. 1 | GLI1⁺ cells are localized adjacent to the base of crypts and are positive for the mesenchymal marker PDGFRA.

a, Sections of duodenal or colonic crypts. Immunohistochemistry for GLI1 (red), E-cadherin (green, top) or ACTA2 (green, bottom) and DAPI (blue) in the duodenum and colon. E-cadherin marks epithelial cells, ACTA2 marks myofibroblasts (mesenchymal cells). b, Cross-sections of duodenal or colonic crypts stained as in a. c, Expression of GLI1 overlaps with that of PDGFRA in mesenchymal cells of Gli1-Cre^{ERT2};lox-STOP-lox-tdTomato⁺; PdgfraEGFP⁺ mice 24 h after tamoxifen injection. Immunohistochemistry:

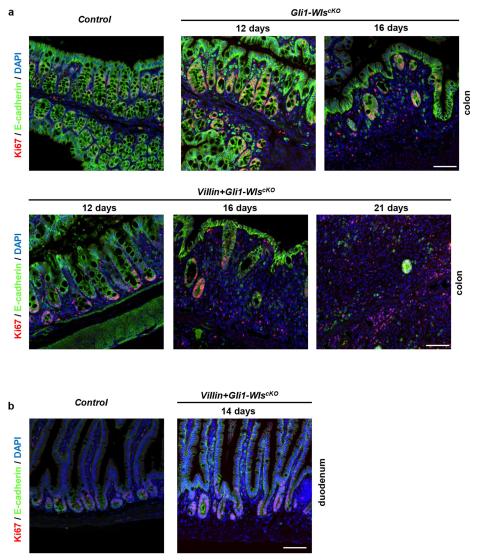
direct detection of fluorescent proteins in frozen sections. tdTomato (red) marks GLI1⁺ cells, nuclear EGFP (green) is expressed in PDGFRA⁺ cells, counterstained with DAPI (blue). d, Intestinal mesenchymal cells isolated from *Gli1-Cre^{ERT2};lox-STOP-lox-tdTomato*⁺;*Pdgfra^{EGFP+}* mice 24 h after tamoxifen injection were cultured in Mesencult medium for 5 days. Immunocytochemistry for EGFP (green) marking PDGFRA, tdTomato (red) for GLI1⁺, DAPI (blue). Full genotype: *Gli1-Cre^{ERT2};lox-STOP-lox-tdTomato*^{Tg/wt};*Pdgfra-EGFP*^{Tg/wt}; Tg indicates transgenic allele. Scale bar, 50 μm; IMCs, intestinal mesenchymal cells.





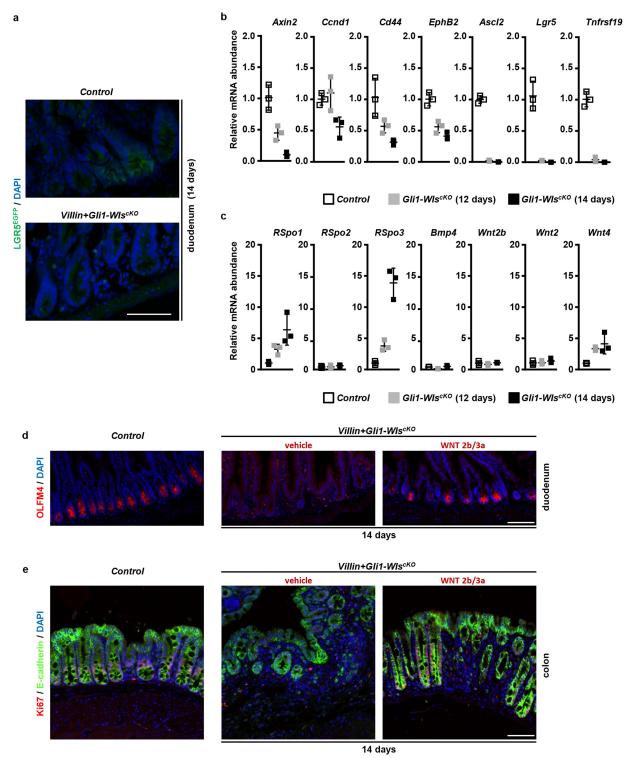
Extended Data Fig. 2 | GLI1 $^+$ cells are persistent in vivo and can act as progenitor cells with potential for tri-lineage differentiation in vitro. a, tdTomato $^+$ cells are retained in the colon mesenchyme of Gli1-Cre^{ERT2};lox-STOP-lox-tdTomato mice 35 days after a single tamoxifen injection. Immunohistochemistry: tdTomato (red) marks GLI1 $^+$ cells or their descendants, E-cadherin (green) denotes the shape of crypts, DAPI (blue) stains nuclei. Full genotype: Gli1-Cre^{ERT2};lox-STOP-lox-

tdTomato^{Tg/w} (Tg indicates transgenic allele). **b**, Sorted GLI1⁺(td Tomato⁺) cells have the capacity to differentiate towards adipocytes (Oil Red O staining), osteoblasts (Alizarin Red S staining) and smooth muscle (ACTA2). **c**, Undifferentiated GLI1⁺ (tdTomato⁺) cells express lower levels of ACTA2 than cells differentiated towards smooth muscle cells. Immunocytochemistry for ACTA2 (green) showing the control (undifferentiated) parallel to **b**. Scale bar, 50 μm.



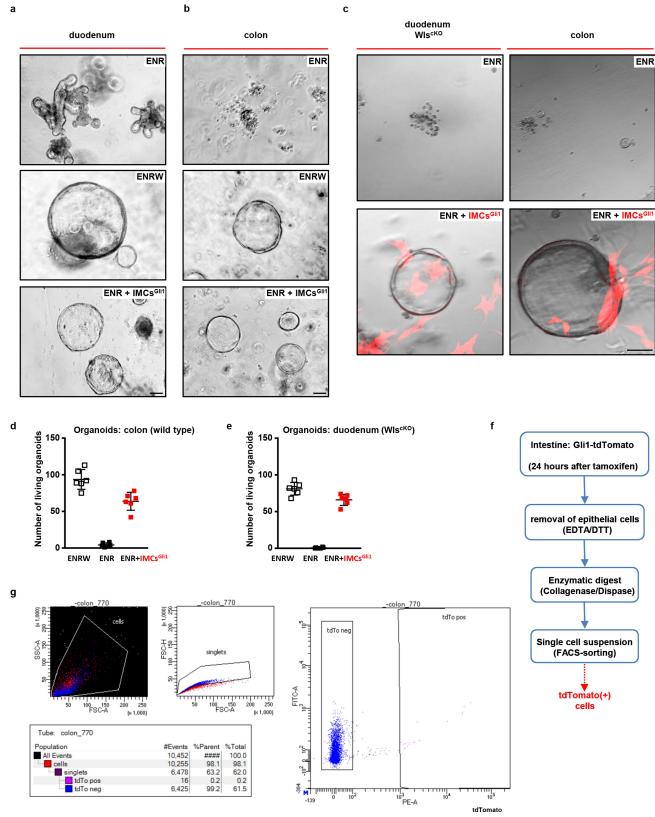
Extended Data Fig. 3 | GLI1 $^+$ cells are essential for the maintenance of the colonic epithelium but dispensable for the normal renewal of the small intestinal (duodenal) epithelium. a, In the colon, blocking Wnt secretion from GLI1 $^+$ cells results in corrupted epithelial morphology in conjunction with loss of proliferation. As no Wnts are secreted from the colonic epithelium, blocking Wnt secretion from both epithelial ($Villin-Cre^{ERT2}$ is active in all epithelial cells) and GLI1 $^+$ cells is

similar to the situation in which Wnt secretion is ablated only in GLI1 $^+$ cells. Immunohistochemistry for Ki67 (red), E-cadherin (green) and DAPI (blue) in colon. Days after Cre induction are indicated. **b**, Normal epithelial morphology of the duodenum in Villin+Gli1-WlscKO animals 12 days after Cre induction. Full genotypes: Gli1-WlscKO (Gli1-CreERT2; Wlsflox/flox), Villin+Gli1-WlscKO (Villin-CreERT2; Gli1-CreERT2; Wlsflox/flox). Scale bar, 50 μm .



Extended Data Fig. 4 | Role of Wnt secretion by GLI1+ cells for the maintenance of intestinal epithelial stem cells differs between duodenum and colon. a, In the duodenum, only simultaneous blocking of Wnt secretion from both the epithelium and GLI1+ cells abrogates the renewal of IES cells marked by LGR5. Immunohistochemistry: LGR5^{EGFP} (green), DAPI (blue), direct detection of fluorescent protein in frozen sections. b, c, In the colon, blocking Wnt secretion from GLI1+ cells results in reduced expression of Wnt target genes and disappearance of IES cell markers (b). Abrogating GLI1-mediated Wnt secretion alters expression of *Rspo1* and *Rspo3* (c); (real-time RT–qPCR, 12 and 14 days after tamoxifen administration, n=3 biologically independent mice, mean \pm s.d.).

d, **e**, Delivery of external Wnts (WNT3A and WNT2B) delays the loss of expression of the stem cell marker OLFM4 in the duodenal crypts of *Villin+Gli1-Wls*^{cKO} mice. Immunohistochemistry: OLFM4 (red), DAPI (blue). **e**, External Wnts (WNT3A and WNT2B) prolong the presence of actively proliferating cells and intact crypts within the colonic epithelium of *Villin+Gli1-Wls*^{cKO} mice. Immunohistochemistry: Ki67 (red), E-cadherin (green, marks epithelial cells), DAPI (blue). Scale bars, 50 μm. Full genotypes: *Gli1-Wls*^{cKO} (*Gli1-Cre*^{ERT2}; *Wls*^{flox/flox}), *Villin+Gli1-Wls*^{cKO} (*Villin-Cre*^{ERT2}; *Gli1-Cre*^{ERT2}; *Wls*^{flox/flox}). Days after Cre induction are indicated.



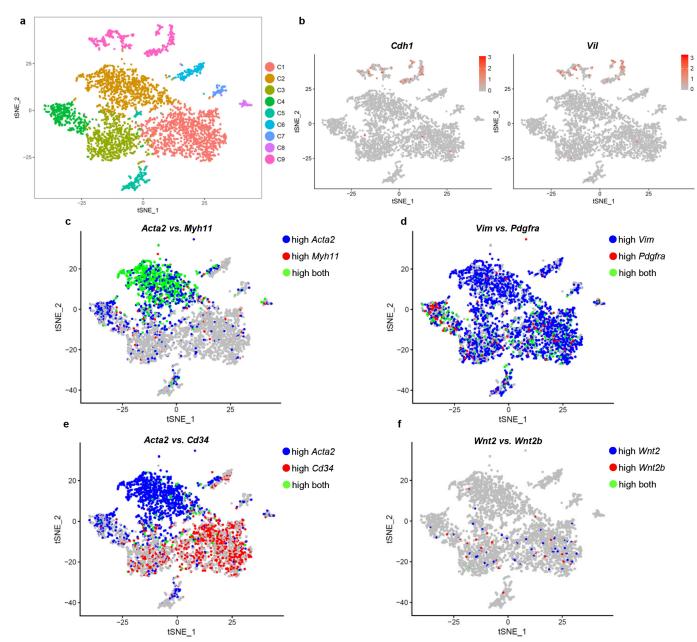
Extended Data Fig. 5 | See next page for caption.



Extended Data Fig. 5 | GLI1⁺ cells restore the growth of intestinal organoids and maintain the intestinal epithelial stem cells in vitro.

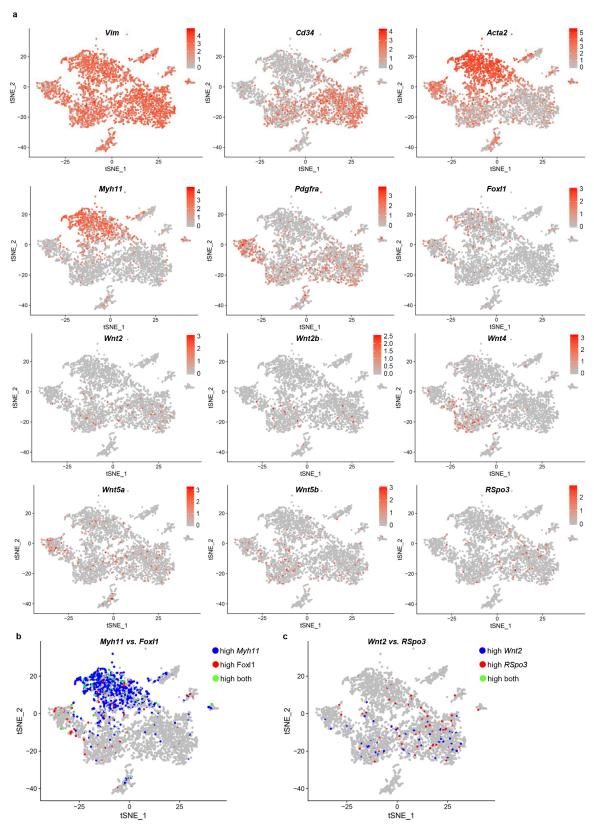
a, GLI1⁺(tdTomato⁺) sorted cells change the morphology of duodenal organoids. Small intestinal (duodenal) organoids were cocultured with sorted GLI1⁺ IMCs. b, c, GLI1⁺(tdTomato⁺) sorted cells drive the growth of colonic organoids by providing Wnt ligands. Colonic organoids were cocultured with or without sorted GLI1⁺ IMCs and grown in medium as indicated. c, GLI1⁺ (tdTomato⁺) cells (red) sustain the growth of colonic organoids and small intestinal (duodenal) organoids with ablated Wnt secretion (Wls^{cKO}). d, e, Coculture with GLI1⁺ cells restored the

growth of colonic and duodenal Wls^{cKO} organoids. The fraction of living organoids after 7 days in culture is shown in the graph. Crypts were seeded at the same initial density. Summarizes data from two independent experiments (n=2 biologically independent experiments, 3 replicates each); mean \pm s.d. Organoids were cultivated in medium containing: ENR = EGF, noggin, R-spondin1; ENRW = EGF, noggin, R-spondin1, WNT3A. IMCs^{GLI1} are sorted GLI1⁺ IMCs. Scale bar, 50 μ m. f, Scheme depicting how GLI1⁺ (tdTomato⁺) cells were sorted. g, Representative gating showing sorting of GLI1⁺ (tdTomato⁺) cells from *Gli1-Cre*^{ERT2};lox-STOP-lox-tdTomato mice.



Extended Data Fig. 6 | GLI1⁺ cells constitute a heterogenous **population.** a, Mesenchymal cells from Villin+Gli1-tdTomato mice ($Villin-Cre^{ERT2}$; $Gli1-Cre^{ERT2}$; lox-STOP-lox-tdTomato) 24 h after tamoxifen injection were isolated by cell sorting (see Methods). Despite the removal of the majority of epithelial cells, some epithelial cells (expressing Villin, E-cadherin and other markers) were identified as a fully distinct cluster C9. This distinct epithelial cluster serves as an internal control for scRNA-seq. Unbiased t-SNE clustering analysis of colon stem cells. Each dot represents an individual cell. 4,464 single cells were successfully profiled; numbers of cells per cluster are: C1 (n=1,267); C2 (n=1,072); C3 (n=723); C4 (n=381); C5 (n=222); C6 (n=159); C7 (n=77); C8

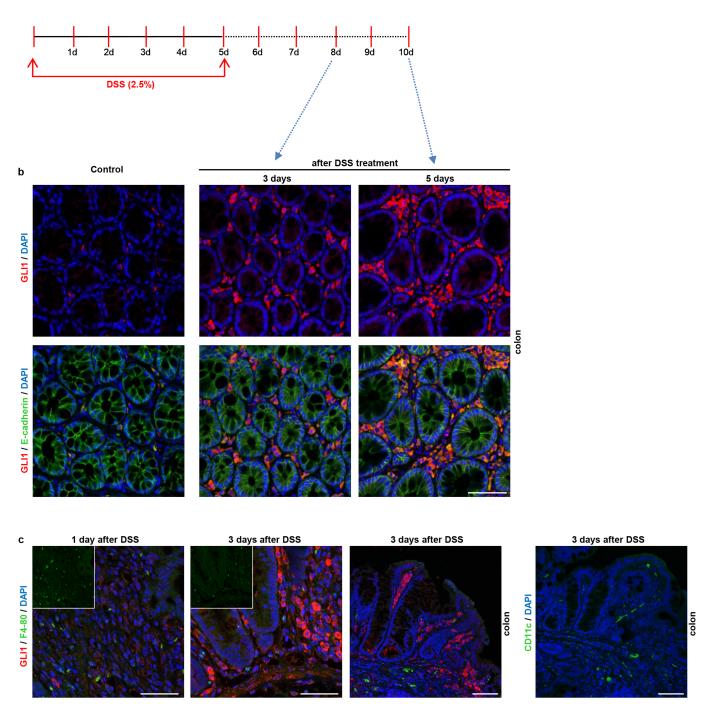
(n=56); C9 (n=507). Colours correspond to unbiased classification via graph-based clustering. **b**, Cdh1 and Vil in epithelial cluster C9. Log-normalized gene expression levels visualized on a t-SNE plot. Each dot represents an individual cell. To show clearly mesenchymal populations, cluster C9 was removed in other panels and figures. **c**-**f**, Co-expression of indicated mesenchymal markers and Wnt ligands simultaneously visualized using t-SNE plot. Each dot represents an individual cell. Blue and red colours indicate individual genes, green colour denotes cells with simultaneously high expression of both genes. (For all panels the cluster of non-mesenchymal cells (C9) was removed.)



Extended Data Fig. 7 | GLI1⁺ cells constitute a heterogeneous population of mesenchymal cells. a, Expression of indicated genes in distinct populations of GLI1⁺ cells. Log-normalized gene expression levels visualized on a t-SNE plot. Each dot represents an individual cell. b, c, Co-expression of *Myh11*, *Foxl1*, *Wnt2* and *RSpo3* simultaneously

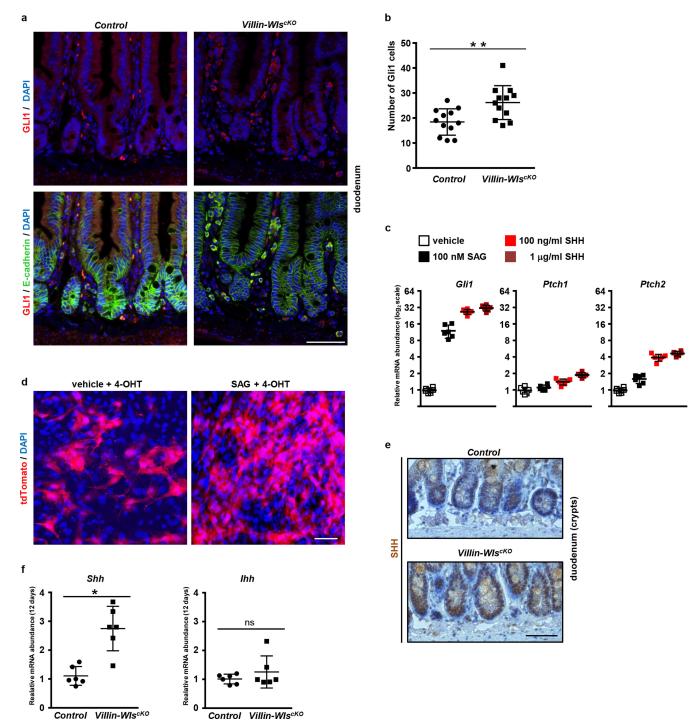
visualized using t-SNE plot. Each dot represents an individual cell. Blue and red colours indicate individual genes, green colour denotes cells with simultaneous high expression of both genes. (For all panels the cluster of non-mesenchymal cells (C9) was removed; for details see Extended Data Fig. 6a.)

а



Extended Data Fig. 8 | GLI1⁺ cells are enriched during epithelial perturbations. a, Scheme of the DSS-induced damage regimen. b, Regeneration of the colonic epithelium during recovery from DSS-mediated damage is associated with an increased number of GLI1⁺ cells. Immunohistochemistry for GLI1 (red), DAPI (blue) and E-cadherin (green) in the colon, 3 or 5 days after the termination of DSS treatment. c, The increase in GLI1⁺ cells during regeneration after DSS-mediated damage does not correspond to the increased number of myeloid cells,

such as macrophages (marked by F4-80) and CD11C-positive cells. Immunohistochemistry for GLI1 (red), and F4-80 (green) or CD11C (green, right) in the colon, 1 or 3 days after termination of DSS treatment. DAPI marks nuclei. Insets depict the individual green channels. The panel with CD11C staining shows a similar area to the panel with GLI1/F4-80 staining (second from right). Owing to antibody incompatibility it was not possible to perform GLI1/CD11C double staining. Scale bar, 50 μm .



Extended Data Fig. 9 | In the duodenum, blocking Wnt secretion from epithelial cells is compensated by an increased number of GLI1⁺ cells that respond to Hedgehog pathway activation. a, The loss of Wnt secretion in the duodenal epithelium is compensated by an increased number of GLI1⁺ cells in the intestinal sub-epithelial layer. Immunohistochemistry for GLI1 (red), DAPI (blue) and/or E-cadherin (green) in the duodenum. $\boldsymbol{b},$ Quantification of GLI1 $^+$ cells in control (n = 4 independent mice) and Villin-Wls^{cKO} (n = 4 independent mice)duodena. For each mouse, three different pictures showing transverse sections were counted. The individual data points show the average number of GLI1⁺ cells per picture; mean \pm s.d. ** $P \le 0.01$, (t-test, one-sided), P = 0.008. c, Intestinal mesenchymal cells (from the duodenum) respond to stimulation of the Hedgehog pathway via recombinant SHH or smoothened agonist (SAG). Changes in the expression levels of Hedgehog target genes Gli1, Ptch1 and Ptch2 (RT-qPCR, n = 2 biologically independent experiments, each 3

replicates, mean \pm s.d.). **d**, Activation of the Hedgehog pathway by 100 nM smoothened agonist (SAG) increased the number of GLI1⁺ cells in (unsorted) intestinal mesenchymal cells from the duodenum. *Gli1-Cre* ERT2; *Jox-STOP-lox-tdTomato* mesenchymal cells were cultured for 5 days with (or without) 100 mM SAG and for the last 12 h with 500 nM 4-hydroxytamoxifen (4-OHT). Immunocytochemistry: tdTomato (red) marks GLI1⁺ cells, DAPI (blue). **e**, Blocking Wnt secretion from the small intestinal epithelium results in elevated levels of SHH. Immunohistochemistry: SHH (brown; DAB), nuclei (haematoxylin). **f**, Relative increase in the expression of *Shh* in the duodenum of *Villin-Wls* mice. IHH is another ligand secreted by cells of the intestinal epithelium; its expression is unchanged (RT-qPCR, 12 days after tamoxifen administration, n = 6 independent animals; mean \pm s.d. $*P \le 0.05$, (one-sided t-test); P(SHH) = 0.002, P(IHH) = 0.2). Scale bar, 50 µm. Full genotypes: $V(\text{Illin-Wls}^{cKO}(\text{Villin-Cre}^{ERT2}; \text{Wls}^{flox/flox})$.



Induction and transcriptional regulation of the co-inhibitory gene module in T cells

Norio Chihara^{1,8}, Asaf Madi^{1,2,8}, Takaaki Kondo¹, Huiyuan Zhang¹, Nandini Acharya¹, Meromit Singer², Jackson Nyman², Nemanja D. Marjanovic², Monika S. Kowalczyk^{2,7}, Chao Wang¹, Sema Kurtulus¹, Travis Law², Yasaman Etminan¹, James Nevin¹, Christopher D. Buckley³, Patrick R. Burkett^{1,4}, Jason D. Buenrostro², Orit Rozenblatt-Rosen², Ana C. Anderson^{1,2,9}*, Aviv Regev^{2,5,6,9}* & Vijay K. Kuchroo^{1,2,9}*

The expression of co-inhibitory receptors, such as CTLA-4 and PD-1, on effector T cells is a key mechanism for ensuring immune homeostasis. Dysregulated expression of co-inhibitory receptors on CD4+ T cells promotes autoimmunity, whereas sustained overexpression on CD8⁺ T cells promotes T cell dysfunction or exhaustion, leading to impaired ability to clear chronic viral infections and diseases such as cancer^{1,2}. Here, using RNA and protein expression profiling at single-cell resolution in mouse cells, we identify a module of co-inhibitory receptors that includes not only several known co-inhibitory receptors (PD-1, TIM-3, LAG-3 and TIGIT) but also many new surface receptors. We functionally validated two new co-inhibitory receptors, activated protein C receptor (PROCR) and podoplanin (PDPN). The module of coinhibitory receptors is co-expressed in both CD4⁺ and CD8⁺ T cells and is part of a larger co-inhibitory gene program that is shared by non-responsive T cells in several physiological contexts and is driven by the immunoregulatory cytokine IL-27. Computational analysis identified the transcription factors PRDM1 and c-MAF as cooperative regulators of the co-inhibitory module, and this was validated experimentally. This molecular circuit underlies the co-expression of co-inhibitory receptors in T cells and identifies regulators of T cell function with the potential to control autoimmunity and tumour immunity.

We used single-cell RNA sequencing (scRNA-seq) to analyse coinhibitory and co-stimulatory receptor expression in 588 CD8⁺ and 316 CD4⁺ tumour-infiltrating lymphocytes (TILs) from B16F10 mouse melanoma³. We found that the expression of *Pdcd1* (also known as PD-1), Tim3 (Havcr2), Lag3, Ctla4, 4-1BB (Tnfrsf9) and Tigit strongly co-vary in CD8⁺ TILs. CD4⁺ TILs showed a similar pattern with the additional co-expression of Icos, Gitr (also known as Tnfrsf18) and Ox40 (Tnfrsf4) (Fig. 1a, top). Single-cell mass cytometry (cytometry by time of flight, CyTOF) confirmed the surface co-expression of these receptors (Fig. 1a, bottom, Supplementary Information 1). The expression of PD-1, LAG-3, TIM-3 and TIGIT was tightly correlated on both CD8⁺ and CD4⁺ TILs (Fig. 1a, bottom). Clustering analysis (t-stochastic neighbourhood embedding (t-SNE)⁴, Methods) showed two groups of CD8⁺ TILs (clusters 1 and 2) (Fig. 1b, Extended Data Fig. 1a, c), with PD-1, LAG-3, TIM-3 and TIGIT mainly expressed in cluster 1 cells (Fig. 1b, Extended Data Fig. 1c), in addition to LILRB4 (Extended Data Fig. 1a) and co-stimulatory receptors of the TNF receptor family, 4-1BB, OX40 and GITR. By contrast, ICOS and CD226 were less restricted to cluster 1 (Extended Data Fig. 1a). We further observed two discrete clusters of CD4⁺ TILs (clusters 3 and 4), with co-expression of PD-1, TIM-3, LAG-3 and TIGIT restricted to cluster 3 (Fig. 1b, Extended Data Fig. 1c).

The co-expression of co-inhibitory receptors on CD8⁺ and CD4⁺ T cells suggests a common trigger. One candidate is IL-27, a hetero-dimeric member of the IL-12 cytokine family that suppresses auto-immunity⁵, induces IL-10-secreting type 1 regulatory T (T_{reg}) cells^{6,7} and induces expression of TIM-3 and PD-L1 on CD4⁺ and CD8⁺ T cells^{8,9}. Activation of CD4⁺ and CD8⁺ T cells in the presence of IL-27 induced the expression of TIM-3, LAG-3 and TIGIT at both the mRNA (Fig. 1c) and protein levels (Extended Data Fig. 2a). mRNA expression of *Tim3* (*Havcr2*), *Lag3* and *Tigit* was reduced in IL-27RA-deficient T cells, whereas *Pdcd1* expression was unaffected by IL-27 in vitro (Fig. 1c, Extended Data Fig. 2a).

CyTOF analysis showed that the loss of IL-27RA resulted in the loss of cells in cluster 1 of CD8+ TILs and cluster 3 of CD4+ TILs (Fig. 1d, $P=5\times10^{-23}$ and 6.8×10^{-7} for CD8+ and CD4+, respectively, hypergeometric test; Extended Data Fig. 1b–d), indicating a key role for IL-27 in driving co-inhibitory receptor co-expression in both CD4+ and CD8+ T cells in vivo. Although PD-1 expression was not dependent on IL-27 in vitro, it was dependent on IL-27RA signalling in vivo. Consistent with the induction of IL-10 by IL-27⁵⁻⁷, we observed reduced IL-10 in IL-27RA-knockout CD8+ TILs (Extended Data Fig. 2b).

scRNA-seq of CD8⁺ and CD4⁺ TILs from wild-type and IL-27RA-knockout mice (Fig. 1e, Extended Data Fig. 3a, b, Methods) revealed distinct clusters of CD8⁺ (cluster 5) and CD4⁺ (cluster 4) TILs that highly expressed the co-inhibitory receptors *Pdcd1*, *Tim3*, *Lag3* and *Tigit*. The expression of these genes was decreased in CD8⁺ TILs from IL-27RA-knockout mice, whereas the expression of only *Tim3* and *Lag3* was decreased in CD4⁺ TILs from IL-27RA-knockout mice (Fig. 1e). Thus, IL-27 drives a module of co-inhibitory receptors that are strongly co-expressed in vivo together with IL-10.

The co-inhibitory receptor module could be part of a larger IL-27-driven inhibitory gene program. We analysed the mRNA profiles of CD4+ and CD8+ T cells stimulated in the presence or absence of IL-27. IL-27 induced similar expression programs in CD4+ and CD8+ T cells (Extended Data Fig. 4a, b). We identified 1,201 genes with IL-27-dependent expression (Methods). We compared the IL-27-driven gene program to the gene signatures for four different states of T cell non-responsiveness: CD8+ T cell exhaustion in both cancer³ and chronic viral infection¹0, and antigen-specific¹¹ and non-specific (anti-CD3 antibody¹²) CD4+ T cell tolerance. We found a significant overlap with all of these signatures (Methods, Extended Data Fig. 4c-f).

Projection of the IL-27 and CD8⁺ cancer T cell exhaustion overlap signature onto the single-cell profiles of CD8⁺ TILs marked a distinct subset of cells (Fig. 2a, panel I). This subset scored highly for the overlap signatures between the IL-27-driven gene program and each of the

¹Evergrande Center for Immunologic Diseases and Ann Romney Center for Neurologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, MA, USA. ²Broad Institute of MIT and Harvard, Cambridge, MA, USA. ³Rheumatology Research Group, Center for Translational Inflammation Research, Queen Elizabeth Hospital, Birmingham, UK. ⁴Pulmonary and Critical Care Division, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. ⁵Howard Hughes Medical Institute, Koch Institute for Integrative Cancer Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁶Department of Biology, Koch Institute and Ludwig Center, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁷Present address: Celsius Therapeutics, Cambridge, MA, USA. ⁸These authors contributed equally: Norio Chihara, Asaf Madi. ⁹These authors jointly supervised this work: Ana C. Anderson, Aviv Regev, Vijay K. Kuchroo. *e-mail: acanderson@partners.org: aregev@broadinstitute.org: vkuchroo@evergrande.hms.harvard.edu

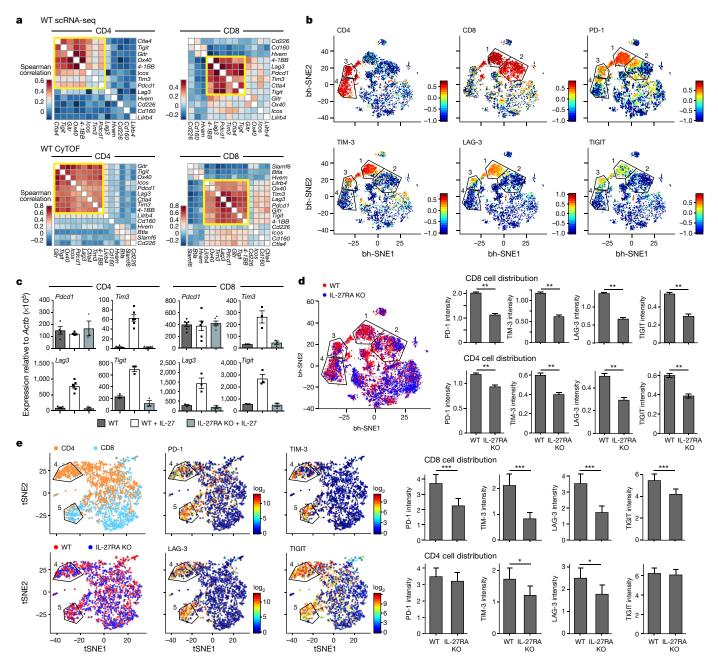


Fig. 1 | Several co-inhibitory receptors are expressed as a module on CD4⁺ and CD8⁺ T cells. a, CD4⁺ and CD8⁺ TILs were obtained from wild-type (WT) mice bearing B16F10 melanoma tumours. Top, co-expression analysis of co-inhibitory and co-stimulatory receptor mRNA expression as determined by scRNA-seq for 316 CD4⁺ and 588 CD8⁺ TILs. Bottom, protein expression by CyTOF for 23,656 CD4⁺ and 36,486 CD8⁺ TILs. Spearman correlation, followed by dendrogram ordering of the matrix using Euclidian distance, is shown. Data are from biologically independent experiments. *Hvem* is also known as *Tnfrsf14*. b, TILs from wild-type mice bearing B16F10 melanoma were analysed using CyTOF with a custom panel of antibodies against co-inhibitory and co-stimulatory cell-surface receptors^{2,24} (Supplementary Table 1). Data were analysed using viSNE. Polygons indicating clusters 1 and 2 (in CD8⁺ T cells), and 3 and 4 (in CD4⁺ T cells) are shown. Individual panels show expression of the indicated markers. c, Naive T cells from either wild-type or IL-27RA-

other three states of T cell non-responsiveness (Fig. 2a, panels II–IV). The transcriptional program induced in IL-27RA-knockout TILs was active in a complimentary subset of TILs (Fig. 2a, panel V, Methods). The control signature from cells stimulated with IL-27 in vitro showed bimodal distribution and by itself did not detect the same population of cells (Fig. 2a, panel VI). From these analyses, we identified a

knockout (KO) mice were stimulated with anti-CD3/CD28 in the presence or absence of IL-27. The indicated expression of co-inhibitory receptors was examined by quantitative PCR (qPCR) at 96 h (CD4) and 72 h (CD8). Data are mean \pm s.e.m. from biologically independent animals. $\bf d$, viSNE plot showing wild-type (red) and IL-27RA-knockout (blue) cells. $\bf e$, scRNA-seq of TILs from mice bearing B16F10 melanoma. Data were analysed using t-SNE. Polygons indicating clusters 4 (in CD4+ T cells, orange) and 5 (in CD8+ T cells, blue) are shown. Individual panels show expression of the indicated markers. Bar graphs show the mean signal intensity for indicated co-inhibitory receptors from: WT (CD4+ (n=849); CD8+ (n=1752)) and IL-27RA-KO (CD4+ (n=628); CD8+ (n=541)) TILs for CyTOF ($\bf d$) or WT (CD4+ (n=707); CD8+ (n=825)) and IL-27RA-KO (CD4+ (n=376); CD8+ (n=394)) TILs for scRNA-seq ($\bf e$). Error bars indicate s.e.m. *P<0.05, **P<0.01, ***P<0.001, two-sided t-test.

co-inhibitory gene module (272 genes) that is shared across several states of T cell non-responsiveness (Supplementary Table 2). Within this module, we identified a set of 57 genes that encode cell-surface receptors and cytokines, including TIM-3, LAG-3, TIGIT and IL-10 (Fig. 2b), which we further stratified by their expression in cancer and chronic viral infections (Fig. 2c). Two surface molecules, PROCR

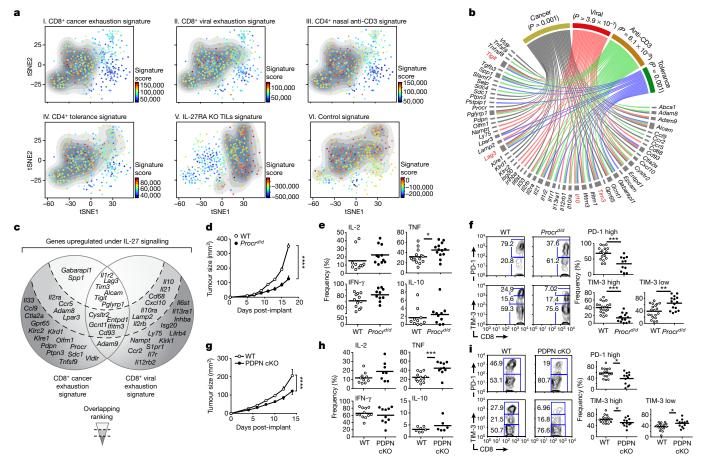


Fig. 2 | The IL-27-induced gene program overlaps with multiple signatures of T cell dysfunction and tolerance. a, Panels I–VI, *t*-SNE plots of the 588 CD8⁺ single-cell TILs (dots) collected from wild-type mice bearing B16F10 melanoma. Cells are coloured by their signature score that reflects the relative average expression of the genes in the overlap of the IL-27-induced gene program with the signatures for each of the indicated states of T cell non-responsiveness. Panel VI is a projection of a signature of the differentially expressed genes between CD8⁺ TILs from wild-type and IL-27RA-knockout mice bearing B16F10 melanoma (Methods). The contour marks the region of highly scored cells based on cells with signature scores above the mean. b, Graphical representation of the overlap of 57 IL-27-induced cell-surface receptors or cytokine genes with genes expressed in different states of T cell non-responsiveness. The width of the grey bars reflects the extent of overlap across states. The significance of the overlap genes between the IL-27-induced state and each

state of T cell non-responsiveness was calculated using Wilcoxon mean rank gene set test (WilcoxGST) and camera. **c**, Graphical representation of the selected overlap genes between the cancer exhaustion and the chronic viral exhaustion signatures. The shaded background reflects the ranking based on the extent of overlap with the T cell states depicted. **d**, **g**, Wild-type (n=8) and $Procr^{d/d}$ (n=7) mice (**d**) or wild-type (n=5) and PDPN cKO (n=5) mice (**g**) were implanted with B16F10 melanoma. Data are mean \pm s.e.m. from three biologically independent experiments. ****P < 0.0001, repeated measures ANOVA, Sidak's multiple comparisons test. **e**, **h**, Summary of flow cytometry data for cytokine production in the indicated CD8⁺ TILs. Data are from biologically independent animals. Horizontal lines denote mean values. **f**, **i**, Left, representative flow cytometry data for TIM-3 and PD-1 expression on the indicated CD8⁺ TILs. Right, summary data. *P < 0.05, **P < 0.01, ***P < 0.001, two-sided t-test (**e**-**i**).

and PDPN, were highly expressed in the setting of cancer (Fig. 2c). Activation of naive $\mathrm{CD4^+}$ and $\mathrm{CD8^+}$ T cells in vitro in the presence of IL-27 induced the expression of PROCR and PDPN (Extended Data Fig. 5a). In vivo, PROCR and PDPN exhibited IL-27-dependent coexpression with PD-1 and TIM-3 on $\mathrm{CD8^+}$ TILs (Extended Data Fig. 5b).

PROCR+CD8+ TILs exhibited an exhausted phenotype, producing less TNF and IL-2 and more IL-10 than PROCR-CD8+ TILs (Extended Data Fig. 5c). The growth of B16F10 melanoma was inhibited in PROCR hypomorph ($Procr^{delta/delta}$, hereafter $Procr^{d/d}$)¹³ mice (Fig. 2d), and $Procr^{d/d}$ CD8+ TILs mice exhibited enhanced production of TNF, but no difference in the production of IL-2, IFN- γ or IL-10 (Fig. 2e). $Procr^{d/d}$ TILs exhibited a decreased frequency of TIM-3^{high} and PD-1^{high} CD8+ T cells, suggesting that PROCR signalling promotes a severely exhausted phenotype in CD8+ T cells ¹⁴ (Fig. 2f). Adoptive transfer of CD8+ T cells that lack PROCR revealed a T cell-specific role for PROCR in constraining tumour growth (Extended Data Fig. 5d).

Although PDPN can limit CD4⁺ T cell survival in inflamed tissues¹⁵, its role in T cell exhaustion is unknown. We observed a significant delay in B16F10 tumour growth in mice with PDPN deficiency in T cells

(PDPN conditional knockout (cKO)) (Fig. 2g). PDPN-deficient CD8 $^+$ TILs exhibited enhanced TNF production but no significant difference in IL-2, IFN- γ or IL-10 (Fig. 2h). The frequency of TIM-3 $^{\rm high}$ and PD-1 $^{\rm high}$ CD8 $^+$ TILs was decreased, indicating a reduced accumulation of T cells with a severely exhausted phenotype in PDPN cKO mice $^{\rm 14}$ (Fig. 2i). Consistent with previous data $^{\rm 15}$, PDPN-deficient PD-1 $^+$ TIM-3 $^+$ CD8 $^+$ TILs had higher expression of IL-7RA, indicating that PDPN may limit the survival of CD8 $^+$ TILs in the tumour microenvironment (Extended Data Fig. 5e, f).

We identified the transcription factor PRDM1 as a candidate regulator of the co-inhibitory module. PRDM1 is induced in vitro by IL-27 in CD4⁺ and CD8⁺ T cells (Extended Data Fig. 6a), is enriched in TILs with high expression of the IL-27 co-inhibitory module (Extended Data Figs. 3c–f, 6b, c, Methods), and is overexpressed in exhausted CD8⁺ TILs ($P\!=\!0.0004$, t-test, Extended Data Fig. 6d). Network analysis based on profiling of naive CD8⁺ T cells from mice with a T cell-specific deletion of PRDM1 (PRDM1 cKO) stimulated with IL-27, showed that PRDM1 regulates several genes in the IL-27 co-inhibitory module (Extended Data Fig. 6e, $P\!=\!2.32\times10^{-12}$;

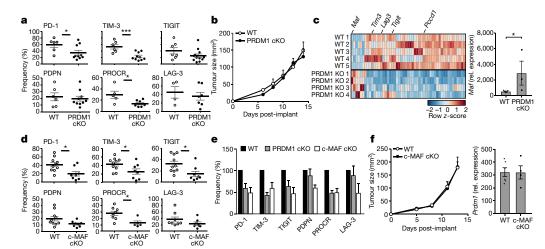


Fig. 3 | PRDM1 and c-MAF individually regulate co-inhibitory receptors on T cells. a, Summary data of co-inhibitory receptor protein expression on CD8+ TILs from wild-type and PRDM1 cKO mice bearing B16F10 melanoma. Data are mean \pm s.e.m. from biologically independent animals. *P < 0.05, ***P < 0.001, two-sided t-test. b, Wild-type (n = 5) and PRDM1 cKO (n = 5) mice were implanted with B16F10 melanoma. Data are mean \pm s.e.m. from three biologically independent experiments. c, Left, gene expression in CD8+ TILs from wild-type and PRDM1 cKO mice bearing B16F10 melanoma was analysed by nCounter codeset (Supplementary Table 3). Differentially expressed genes are shown as a heat map. Right, expression of Maf in CD8+ TILs from wild-type and PRDM1 cKO mice as determined by qPCR. Data are mean \pm s.e.m. from

biologically independent animals. *P = 0.03, two-sided t-test. **d**, Summary data of co-inhibitory receptor protein expression on CD8⁺ TILs from wild-type and c-MAF cKO. Data are mean \pm s.e.m. from biologically independent animals. *P < 0.05, two-sided t-test. **e**, Frequency of co-inhibitory receptor expression of PRDM1 cKO (grey bar) and c-MAF cKO (open bar) CD8⁺ TILs relative to wild type (filled bar). Data are mean \pm s.e.m., calculated based on data from **a** and **d** with wild-type set to 100%. **f**, Left, wild-type (n = 8) and c-MAF cKO (n = 5) mice were implanted with B16F10 melanoma. Data are mean \pm s.e.m. from two biologically independent experiments. Right, expression of Prdm1 in CD8⁺ TILs from wild-type and c-MAF cKO mice as determined by qPCR (expression levels relative to Actb).

hypergeometric test; Methods). This was further supported by PRDM1 chromatin immunoprecipitation followed by sequencing (ChIP–seq) data 16 ($P\!=\!2.9\times10^{-8}$, hypergeometric test; Extended Data Fig. 6e, Methods).

CD8⁺ TILs from B16F10 tumour-bearing PRDM1 cKO mice expressed lower levels of TIM-3, PD-1 and PROCR (Fig. 3a); however, there was no difference in tumour growth compared to wildtype controls (Fig. 3b), indicating that the reduction of co-inhibitory receptor expression in PRDM1 cKO mice was insufficient to promote effective anti-tumour immunity. We therefore examined whether other transcription factors may regulate the co-inhibitory module and compensate for the absence of PRDM1. We analysed CD8⁺ TILs from PRDM1 cKO mice for the expression of genes from the IL-27-driven gene signature and the signature for exhausted CD8⁺ TILs (Methods, Supplementary Table 3). We found that only a few genes were upregulated in PRDM1 cKO CD8⁺ T cells, including one transcription factor, c-MAF (*P* < 0.05; Fig. 3c). Indeed, c-MAF is induced by IL-27, is coexpressed with PRDM1 in T cells after IL-27 stimulation (Extended Data Fig. 6a), and can regulate IL-10 expression¹⁷ and T cell exhaustion¹⁸. In addition, many genes (226 genes, $P = 5.34 \times 10^{-5}$, hypergeometric test) in the co-inhibitory gene module have a binding motif and a reported binding event for c-MAF within their promoter regions¹⁹.

CD8⁺ TILs from c-MAF cKO mice exhibited decreased expression of several co-inhibitory receptors (Fig. 3d). PRDM1 and c-MAF each affected co-inhibitory receptor expression only partially (Fig. 3e). As in the PRDM1 cKO mice, c-MAF cKO mice did not show any differences in tumour growth relative to controls (Fig. 3f). Notably, PRDM1 expression in c-MAF cKO TILs was similar to that in wild-type TILs, indicating that PRDM1 might drive the expression of the co-inhibitory gene module in the absence of c-MAF.

We addressed whether PRDM1 and c-MAF could act cooperatively to regulate co-inhibitory receptor expression. We found no evidence for a physical interaction between PRDM1 and c-MAF (data not shown); we therefore examined whether they shared targets. We combined the network analysis for PRDM1 (Extended Data Fig. 6e) with c-MAF ChIP–seq data¹⁹ and c-MAF targets (Methods). We observed 121 genes in the co-inhibitory module that are affected (RNA-seq) or

have a direct binding event (ChIP-seq) for both PRDM1 and c-MAF (Fig. 4a), but that are not affected in either individual knockout. This is consistent, among other possibilities, with compensatory (for example, 'OR' gate logic) regulation²⁰. Examination of ATAC-seq (assay for transposase-accessible chromatin using sequencing)^{21,22} and ChIP-seq data for PD-1, TIM-3, LAG-3 and TIGIT shows that PRDM1 and c-MAF can bind both overlapping and non-overlapping sites in the loci of these receptors and can synergistically trans-activate TIM-3 expression (Extended Data Fig. 7).

Mice with a T cell-specific deletion in both PRDM1 and c-MAF (PRDM1/c-MAF conditional double-knockout (cDKO)) showed normal development of CD4⁺ and CD8⁺ T cells in terms of frequency and expression of memory or activation markers, although the frequency of FOXP3⁺ T_{reg} cells was increased (Extended Data Fig. 8a). CD4⁺ and CD8⁺ TILs from cDKO mice bearing B16F10 melanomas exhibited a near absence of PD-1, TIM-3, LAG-3, TIGIT, PDPN and PROCR expression (Fig. 4b, Extended Data Fig. 8b). Moreover, cDKO CD8⁺ TILs exhibited enhanced IL-2 and TNF production (Extended Data Fig. 8c). In contrast to singly deficient mice, cDKO mice showed significant control of B16F10 tumour growth despite the increased frequency of Treg cells (Fig. 4c). We addressed whether PRDM1 and c-MAF have a cell-intrinsic role in CD8⁺ and CD4⁺ T cells in controlling tumour growth by using an adoptive transfer model. Although CD8⁺ T cells from cDKO were able to inhibit tumour growth with decreased expression of co-inhibitory molecules, these effects were stronger when PRDM1 and c-MAF were lacking in both CD4⁺ and CD8⁺ T cells (Fig. 4d, Extended Data Fig. 8d). We examined the roles of PRDM-1 and c-MAF in tumour antigen-specific T cell responses using the MC38-OVA tumour model. We observed a significant reduction in tumour growth in mice receiving cDKO T cells as compared to mice receiving wild-type T cells (Extended Data Fig. 8e). We also observed an increase in ovalbumin (OVA)-specific T cells in the tumour draining lymph nodes and in OVA-specific IFN-γ- and TNF-producing CD8⁺ T cells in both the tumour infiltrate and the periphery in mice receiving double-knockout T cells (Fig. 4e, f, Extended Data Fig. 8f). Lastly, we observed an increase in CD8⁺Ki67⁺ T cells in the periphery of mice receiving double-knockout T cells (Fig. 4f).

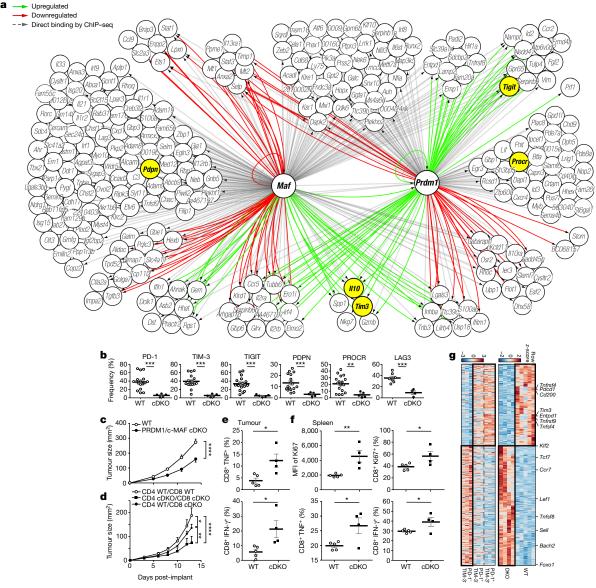


Fig. 4 | PRDM1 and c-MAF together regulate a co-inhibitory gene module that determines anti-tumour immunity. a, Network model based on coupling RNA-seq gene expression data of naive CD8⁺ T cells from PRDM1 cKO or c-MAF cKO mice stimulated in the presence of IL-27 and PRDM1 and c-MAF ChIP-seq data. Upregulated genes (green arrows), downregulated genes (red arrows), and c-MAF or PRDM1 binding events (grey arrows) are shown. **b**, Summary data of indicated co-inhibitory receptors expression on CD8+ TILs from wild-type and PRDM1/c-MAF cDKO mice bearing B16F10 melanoma. Data are mean \pm s.e.m. from biologically independent animals. **P < 0.01, ***P < 0.001, two-sided *t*-test. **c**, Wild-type (n = 15) and cDKO (n = 8) mice were implanted with B16F10 melanoma. Data are from three biologically independent experiments. d, CD4+ or CD8+ T cells sorted from cDKO mice or littermate controls were transferred into RAG1-knockout mice at a 2:1 CD4:CD8 ratio, followed by subcutaneous injection of B16-OVA (n = 5, each condition). Data are representative of three biologically

We tested for non-additive effects between PRDM1 and c-MAF by using a binomial generalized linear model to compare the effect of single knockouts to the cDKO, and found that 149 out of 940 differentially expressed genes (adjusted P < 0.05, likelihood ratio test and false discovery rate (FDR) correction) between wild-type and cDKO CD8⁺ TILs have non-additive (that is, synergistic) effects (Extended Data Fig. 9, Methods).

Examination of the transcriptional signatures of cDKO CD8 $^+$ TILs showed significant overlap with those of CD8 $^+$ TIM-3 $^-$ PD-1 $^-$ TILs

independent experiments. **c**, **d**, Tumour size. Data are mean \pm s.e.m. $^*P < 0.05, ^{**}P < 0.01, ^{****}P < 0.0001, repeated measures ANOVA, Sidak's multiple comparisons test.$ **e**,**f**, T cells were obtained from RAG1-knockout mice that received an adoptive transfer of CD4+ and CD8+ T cells from wild-type or cDKO mice (2:1 ratio of CD4*CD8) followed by subcutaneous injection of MC38-OVA (Extended Data Fig. 8e).**e** $, The frequency of IFN-<math display="inline">\gamma$ and TNF CD8+ TILs after OVA-peptide stimulation. **f**, The frequency and expression of Ki67+ cells on splenocytes (top), and the frequency of IFN- γ and TNF CD8+ splenocytes (bottom) after OVA-peptide stimulation. Data are mean \pm s.e.m. from biologically independent animals. $^*P < 0.05, ^**P < 0.01$, two-sided t-test. **g**, Nine hundred and forty differentially expressed genes between CD8+ TILs from wild-type and cDKO mice bearing B16F10 melanoma. Adjusted P < 0.05, likelihood ratio test and FDR correction (top) and their corresponding expression pattern in PD-1+TIM-3+CD8+, PD-1+TIM-3-CD8+, and PD-1-TIM-3-CD8+ TILs.

(Fig. 4g, $P=2.8\times10^{-7}$, one-sample Kolmogorov–Smirnov test; Extended Data Fig. 10a–c), suggesting that the loss of both c-MAF and PRDM1 increases the proportion of non-exhausted CD8⁺ effectors that exist normally in tumours. We scored the individual scRNA-seq profiles of CD8⁺ TILs for the cDKO 940 gene signature and found that the expression of the cDKO gene signature and the co-inhibitory gene module signature mark mutually exclusive populations of TILs (Extended Data Fig. 10e). The cDKO signature showed significant overlap with PD-1⁺CXCR5⁺CD8⁺ T cells,

which may represent precursors for functional effectors in chronic lymphocytic choriomeningitis virus (LCMV) infection²³ (Extended Data Fig. 10d, e, $P=1\times 10^{-13}$, one-sample Kolmogorov–Smirnov test). Furthermore, the IL-27RA-knockout TIL signature also showed significant overlap with this PD-1+CXCR5+CD8+ T cell signature ($P<2.2\times 10^{-16}$, one-sample Kolmogorov–Smirnov test; Fig. 2a, Extended Data Fig. 10e). Collectively, our data indicate that the loss of c-MAF and PRDM1 preferentially results in loss of the co-inhibitory gene module expression and acquisition of a more responsive effector T cell state.

In conclusion, we identified a co-inhibitory gene module, which is expressed in several settings of both $\mathrm{CD4^+}$ and $\mathrm{CD8^+}$ T cell non-responsiveness, along with its transcriptional regulators. The discovery of this module provides a basis for the identification of novel co-inhibitory and co-stimulatory receptors that may have an important role in T cell regulation.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0206-z.

Received: 16 August 2016; Accepted: 27 April 2018; Published online: 13 June 2018

- Wherry, E. J. & Kurachi, M. Molecular and cellular insights into T cell exhaustion. Nat. Rev. Immunol. 15, 486–499 (2015).
- Anderson, A. C., Joller, N. & Kuchroo, V. K. Lag-3, Tim-3, and TIGIT: co-inhibitory receptors with specialized functions in immune regulation. *Immunity* 44, 989–1004 (2016).
- Singer, M. et al. A distinct gene module for dysfunction uncoupled from activation in tumor-infiltrating T cells. Cell 166, 1500–1511 (2016).
- Maaten, L. H. G. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008).
- Fitzgerald, D. C. et al. Suppression of autoimmune inflammation of the central nervous system by interleukin 10 secreted by interleukin 27-stimulated T cells. Nat. Immunol. 8, 1372–1379 (2007).
- Awasthi, A. et al. A dominant function for interleukin 27 in generating interleukin 10-producing anti-inflammatory T cells. *Nat. Immunol.* 8, 1380–1389 (2007).
- Stumhofer, J. S. et al. Interleukins 27 and 6 induce STAT3-mediated T cell production of interleukin 10. Nat. Immunol. 8, 1363–1371 (2007).
- Zhu, C. et al. An IL-27/NFIL3 signalling axis drives Tim-3 and IL-10 expression and T-cell dysfunction. Nat. Commun. 6, 6072 (2015).
- Hirahara, K. et al. Interleukin-27 priming of T cells controls IL-17 production in trans via induction of the ligand PD-L1. *Immunity* 36, 1017–1030 (2012).
- Doering, T. A. et al. Network analysis reveals centrally connected genes and pathways involved in CD8+T cell exhaustion versus memory. *Immunity* 37, 1130–1144 (2012).
- Burton, B. R. et al. Sequential transcriptional changes dictate safe and effective antigen-specific immunotherapy. *Nat. Commun.* 5, 4741 (2014).
- Mayo, L. et al. IL-10-dependent Tr1 cells attenuate astrocyte activation and ameliorate chronic central nervous system inflammation. *Brain* 139, 1939–1957 (2016).

- Castellino, F. J. et al. Mice with a severe deficiency of the endothelial protein C receptor gene develop, survive, and reproduce normally, and do not present with enhanced arterial thrombosis after challenge. *Thromb. Haemost.* 88, 462–472 (2002).
- Sakuishi, K. et al. Targeting Tim-3 and PD-1 pathways to reverse T cell exhaustion and restore anti-tumor immunity. J. Exp. Med. 207, 2187–2194 (2010).
- Peters, A. et al. Podoplanin negatively regulates CD4+ effector T cell responses. J. Clin. Invest. 125, 129–140 (2015).
- Mackay, L. K. et al. Hobit and Blimp 1 instruct a universal transcriptional program of tissue residency in lymphocytes. Science 352, 459–463 (2016).
- Apetoh, L. et al. The aryl hydrocarbon receptor interacts with c-Maf to promote the differentiation of type 1 regulatory T cells induced by IL-27. Nat. Immunol. 11, 854–861 (2010).
- Giordano, M. et al. Molecular profiling of CD8 T cells in autochthonous melanoma identifies Maf as driver of exhaustion. EMBO J. 34, 2042–2058 (2015).
- Ciofani, M. et al. A validated regulatory network for Th17 cell specification. Cell 151, 289–303 (2012).
- Capaldi, A. P. et al. Structure and function of a transcriptional network activated by the MAPK Hog1. Nat. Genet. 40, 1300–1306 (2008).
- Karwacz, K. et al. Critical role of IRF1 and BATF in forming chromatin landscape during type 1 regulatory cell differentiation. Nat. Immunol. 18, 412–421 (2017).
- Sen, D. R. et al. The epigenetic landscape of T cell exhaustion. Science 354, 1165–1169 (2016).
- Im, S. J. et al. Defining CD8+T cells that provide the proliferative burst after PD-1 therapy. Nature 537, 417–421 (2016).
- 24. Chen, L. & Flies, D. B. Molecular mechanisms of T cell co-stimulation and co-inhibition. *Nat. Rev. Immunol.* **13**, 227–242 (2013).

Acknowledgements We thank M. Collins for discussions, D. Kozoriz, J. Xia and Z. Chen for technical advice, S. Riesenfeld for computational advice, N. Paul and J. Keegan for CyTOF, and L. Gaffney for artwork. This work was supported by grants from the National Institutes of Health, the American Cancer Society, the Melanoma Research Alliance, the Klarman Cell Observatory at the Broad Institute, and the Howard Hughes Medical Institute.

Author contributions N.C., A.M., P.R.B., A.C.A., O.R.-R., A.R. and V.K.K. designed the experiment; N.C., A.M., S.K., J.N., C.D.B., P.R.B., J.D.B. and A.R. developed analytical tools; N.C., A.M., T.K., N.A., J.N., N.D.M., M.S.K., C.W., H.Z., T.L., Y.E. and P.R.B. performed experiments; A.M. and M.S. performed computational analysis. N.C. and A.M. wrote the original draft of the paper and P.R.B., A.C.A., A.R. and V.K.K. reviewed and edited the paper; A.C.A., A.R. and V.K.K. supervised the project.

Competing interests A.C.A. is a member of the SAB for Potenza Therapeutics and Tizona Therapeutics. V.K.K. has an ownership interest and is a member of the SAB for Potenza Therapeutics and Tizona Therapeutics. A.C.A.'s and V.K.K.'s interests were reviewed and managed by the Brigham and Women's Hospital and Partners Healthcare in accordance with their conflict of interest policies. A.R. is an SAB member for Thermo Fisher and Syros Pharmaceuticals and is a consultant for Driver Group.

Additional information

 $\begin{tabular}{ll} \textbf{Extended data} is available for this paper at https://doi.org/10.1038/s41586-018-0206-z. \end{tabular}$

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0206-z.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.C.A. or A.R. or V.K.K.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mice. C57BL/6 wild-type, IL-27RA-knockout, and *Prdm1*^{fl/fl} mice were obtained from the Jackson Laboratory. $Maf^{fl/fl}$, $Pdpn^{fl/fl}$ mice and $Procr^{d/d}$ mice were previously described^{13,15,25}. $Pdpn^{fl/fl}$ mice were initially obtained from C. Buckley and crossed to CD4-Cre mice to obtain conditional deletion in T cells. CD4-Cre mice were purchased from Taconic. $Prdm1^{fl/fl}$ and $Maf^{fl/fl}$ mice were crossed to CD4-Cre mice to generate doubly deficient T cell conditional knockout mice. All experiments were performed in accordance with the guidelines outlined by the Harvard Medical Area Standing Committee on Animals.

Tumour experiments. B16F10 melanoma cells (ATCC) (5×10^5) were implanted into the right flank of C57BL/6 mice. Tumour size was measured in two dimensions using a caliper. TILs were isolated by dissociating tumour tissue in the presence of 2.5 mg ml $^{-1}$ collagenase D for 20 min before centrifugation on a discontinuous Percoll gradient (GE Healthcare). Isolated cells were then used in various assays of T cell function. For antigen specific analysis, we applied adoptive transfer tumour experiments using T cells from PRDM1/c-MAF cDKO mice, CD4+ or CD8+ T cells sorted from cDKO mice or littermate controls were transferred into RAG1knockout mice at a 2:1 ratio (CD4: 1 million per mouse and CD8: 0.5 million per mouse) 2 days before subcutaneous injection of B16-OVA or MC38-OVA tumour. B16-OVA was a gift from K. Wucherpfennig, and MC38-OVA was a gift from M. Smyth. For adoptive transfer tumour experiments using T cells from *Procr*^{d/d} mice, CD4 $^+$ T cells from wild-type and CD8 $^+$ T cells from wild-type or $\mathit{Procr}^{d/d}$ mice were isolated by cell sorting (BD FACS Aria) and transferred into RAGdeficient recipient mice at a 2:1 ratio (WT CD4+: 1 million per mouse and WT or $\mathit{Procr}^{d/d}\,\mathsf{CD8}^{\mp}$: 0.5 million per mouse) 2 days before tumour implant. Although we did not use blinding or randomization experimental approaches, at least five animals of target gene knockout and control mice were used to adequately power biological validation experiments throughout the article. All mice used are C57BL/6 background, both male and female, 6-12 weeks of age, 15-25 g. Each experiment was performed using age- and sex-matched controls (Supplementary Table 5). CyTOF. Antibodies were labelled using MaxPar Metal Labelling Kits (DVS) by The Longwood Medical Area CyTOF Antibody Resource and Core. In some experiments, TILs were enriched using Dynabeads FlowComp Mouse Pan T (CD90.2) Kit (Invitrogen). Cells were washed and resuspended in CyTOF PBS (PBS plus 0.05% sodium azide and 0.5% BSA) and stained viability marker Rhodium (DVS) following the cocktail of antibodies against cell-surface molecules for 30 min. Cells were washed again and resuspended in CyTOF PBS with 4% paraformaldehyde. After 10 min fixation, cells were washed and barcoded with Cell-ID intercalators (DVS). Before analysis, cells were resuspended in water with beads and loaded to the CyTOF Mass Cytometer (DVS). CyTOF data were recorded in dual-count according to Fluidigm's recommended settings that calibrated on the fly, combining pulse-count and intensity information. Data obtained as mass peaks for the channels are processed according to cell event selection criteria. These criteria include cell viability selection (Pt195), single-cell selection (Intercalator-Ir), and barcoding selection (Pt194 and Pt198) to identify single-cell events from wild-type and knockout TILs for further analysis.

To obtain clusters of cells similar in their protein expression patterns, cells were clustered using k-means algorithm. Optimal cluster number was estimated using the within groups sum of squared error (SSE) plot followed by gap statistics with bootstrapping and first standard error max method. These methods suggested 9 clusters as optimal in the multidimensional space. Applying k-means clustering with k=9 on our CyTOF data, resulted in clear distinction between cluster 1 and 2 of the CD8⁺ TILs and cluster 3 and 4 of the CD4⁺ TILs. This separation could be further visualized by two-dimensional nonlinear embedding of the protein expression profiles using t-SNE⁴. The t-SNE plot can then be overlaid by k-means clustering results to reflect a non-biased approach to the clusters or with intensity of the different markers.

Flow cytometry. Single-cell suspensions were stained with antibodies against CD4 (RM4-5), CD8 (53-6.7), PD-1 (RMP1-30), LAG-3 (C9B7W), TIGIT (GIGD7), TIM-3 (5D12), PROCR (eBio1560) and PDPN (8.1.1.) obtained from BioLegend. Fixable viability dye eF506 (eBioscience) was used to exclude dead cells. For intracytoplasmic cytokine (ICC) staining, cells were stimulated with phorbol myristate acetate (50 ng ml $^{-1}$) and ionomycin (1 µg ml $^{-1}$) or with OVA 323-339 peptide for antigen specific experiments. Permeabilized cells were then stained with antibodies against IL-2, TNF, IFN- γ or IL-10. All data were collected on a BD LSR II (BD Biosciences) and analysed with FlowJo software (Tree Star). In brief, data were analysed by FSC/SSC gates of starting cell population (the gating strategy is exemplified in Supplementary Fig. 1). Positive gates were set based on fluorescence minus one (FMO) controls in each setting for cell surface molecules and based on unstimulated sample for ICC staining.

In vitro T cell differentiation. CD4⁺ and CD8⁺ T cells were purified from spleen and lymph nodes using anti-CD4 microbeads and anti-CD8a microbeads (Miltenyi Biotech) then stained in PBS with 0.5% BSA for 15 min on ice with anti-CD4, anti-CD8, anti-CD62L, and anti-CD44 antibodies (all from Biolegend).

Naive CD4⁺ or CD8⁺ CD62L^{high}CD44^{low} T cells were sorted using the BD FACSAria cell sorter. Sorted cells were activated with plate-bound anti-CD3 (2 μg ml $^{-1}$ for CD4 and 1 μg ml $^{-1}$ for CD8) and anti-CD28 (2 μg ml $^{-1}$) in the presence of recombinant mouse IL-27 (25 ng ml $^{-1}$) (eBioscience). Cells were collected at various time points for RNA, intracellular cytokine staining and flow cytometry. ${\bf qPCR}$. Total RNA was extracted using RNeasy columns (Qiagen). Reverse transcription of mRNA was performed in a thermal cycler (Bio-Rad) using iScript cDNA Synthesis Kit (Bio-Rad). qPCR was performed in the Vii7 Real-Time PCR system (Applied Biosystems) using the primers for Taqman gene expression (Applied Biosystems). Data were normalized to the expression of Actb.

Nanostring RNA analysis, expression profiling of TILs. We analysed gene expression in CD8 $^+$ TILs from PRDM1 or c-MAF cKO mice bearing B16F10 melanoma collected on day 14 after tumour implantation, using a custom nanostring codeset of 397 genes representing both the IL-27-driven gene signature (245 genes) and the dysfunctional CD8 $^+$ TIL gene signature (245 genes) (Supplementary Table 3). Expression values were normalized by first adjusting each sample based on its relative value to all samples. This was followed by subtracting the calculated background (mean.2sd) from each sample with additional normalization by house-keeping geometric mean, in which housekeeping genes were defined as: Hprt, Gapdh, Actb and Tubb5. Differentially expressed genes were defined using the function that fits multiple linear models from the Bioconductor package limma in R^{26} with P < 0.05.

Microarray processing and analysis. Naive CD4⁺ and CD8⁺ T cells were isolated from wild-type or IL-27RA-knockout mice, and differentiated in vitro with or without IL-27. Cells were collected at 72 h for CD8⁺ and 96 h for CD4⁺, and Affymetrix GeneChip Mouse Genome 430 2.0 Arrays were used to measure the resulting mRNA levels at these time points. Individual CEL files were RMA normalized and merged to an expression matrix using the ExpressionFileCreator of GenePattern with default parameters²⁷. Gene-specific intensities were then computed by taking for each gene j and sample i the maximal probe value observed for that gene. Samples were then transferred to log-space by taking \log_2 (intensity).

Differentially expressed genes were annotated as genes with FDR-corrected ANOVA <0.05 computed between the CD4 with or without IL-27 stimulation (CD4+ IL-27 and T helper type 0 (TH0)) subpopulations (1,202 genes). A total of 468 genes were differentially expressed between wild-type CD8+ T cells stimulated in the presence or absence of IL-27 (P<0.05). Two hundred and thirty-four genes were shared between these two differentially expressed gene lists ($P=2.25\times10^{-157}$, hypergeometric test, background =16,618 (union of genes expressed)). A list of 972 cell surface/cytokines genes of interest that include: cytokines, adhesion, aggregation, chemotaxis and other cell-surface molecules (Supplementary Table 4) composed using Gene Ontology (GO) annotation in Biomart was used to generate the gene subset in Fig. 2b and c.

RNA-seq gene expression profiling of tumour infiltrating cells. Tumour-infiltrating CD8+ T cells were isolated from wild-type, IL-27RA-knockout, PRDM1 cKO, c-MAF cKO, and PRDM1/c-MAF cDKO tumour-bearing mice via FACS sorting on a FACSAria (BD Biosciences). Tumour-infiltrating CD8+ T cells were processed using an adaptation of the SMART-Seq 2 protocol²⁸, using 5 μ l of lysate from bulk CD8+ T cells as the input for each sample during RNA cleanup via SPRI beads (approximately 2,000 cells lysed on average in RLT).

RNA-seq reads were aligned using Tophat²⁹ (mm9) and RSEM-based quantification³⁰ using known transcripts (mm9), followed by further processing using the Bioconductor package DESeq in R³¹. The data were normalized using TMM normalization. The TMM method estimates scale factors between samples that can be incorporated into currently used statistical methods for DE analysis. Post-processing and statistical analysis was carried out in R³⁰. Differentially expressed genes were defined using the differential expression pipeline on the raw counts with a single call to the function DESeq (adjusted P < 0.1). Heat map figures were generated using pheatmap package (https://CRAN.R-project.org/package=pheatmap).

scRNA-seq. CD4+ and CD8+ TILs from wild-type or IL-27RA-knockout mice bearing B16 melanomas were sorted into 96-well plates with 5 μ l lysis buffer comprised of buffer TCL (Qiagen) plus 1% 2-mercaptoethanol (Sigma). Plates were then spun down for 1 min at 3,000 r.p.m. and immediately frozen at $-80\,^{\circ}\text{C}$. Cells were thawed and RNA was isolated with 2.2× RNAClean SPRI beads (Beckman Coulter Genomics) without final elution 32 . The beads were then air-dried and processed immediately for cDNA synthesis. Samples were then processed using the Smart-seq2 protocol 33 , with minor modifications applied to the reverse transcription step (M.S.K. and A.R., in preparation). This was followed by making a 25 μ l reaction mix for each PCR and performing 21 cycles for cDNA amplification. Then 0.25 ng cDNA from each cell and 0.25 of the standard Illumina NexteraXT reaction volume were used in both the tagmentation and final PCR amplification steps. Finally, libraries were pooled and sequenced (50 × 25 paired-end reads) using a single kit on the NextSeq500 5 instrument. All CD4+ TIL (wild-type and IL-27RA-knockout) scRNA-seq data were generated as part of this study.

CD8⁺ TIL single-cell data include wild-type CD8⁺ TIL data from Singer et al.³ and wild-type and IL-27RA-knockout CD8⁺ single-cell data generated as part of this study. **scRNA-seq data preprocessing and expression.** Initial preprocessing was performed as previously described³. In brief, paired reads were mapped to mouse annotation mm10 using Bowtie³⁴ (allowing a maximum of one mismatch in seed alignment, and suppressing reads that had more than 10 valid alignments) and TPMs were computed using RSEM³⁰, and $\log_2(\text{TPM}+1)$ values were used for subsequent analyses.

Next, we filtered out low-quality cells and cell doublets, maintaining for subsequent analysis the cells that had (1) 1,000–4,000 detected genes (defined by at least one mapped read), (2) at least 200,000 reads mapped to the transcriptome, and (3) at least 50% of the reads mapped to the transcriptome, ending with a total of 707 CD4+ and 825 CD8+ wild-type TILs and 376 CD4+ and 394 CD8+ IL-27RA-knockout TILs. We restricted the genes considered in subsequent analyses to be the genes expressed at $\log_2(\text{TPM}+1) \geq 2$ in at least 20% of the cells.

After removal of low-quality cells, the data were normalized using quantile normalization followed by principal component analysis (PCA). Principal components 1-10 were chosen for subsequent analysis owing to a drop in the proportion of variance explained following principal component 10. We used t-SNE 4 to visualize single-cells in a two-dimensional nonlinear embedding.

scRNA-seq clustering and differential expression analysis. For the coupled dataset of wild-type and IL-27RA-knockout TILs, we followed the analysis previously described³⁵. We performed batch correction using ComBat³⁶ and the batchcorrected expression matrix was then reduced using PCA; principal components 1-13 were chosen for subsequent analysis owing to a drop in the proportion of variance explained following principal component 13. Next, we cluster the cells based on their principal component scores using the Louvain-Jaccard method using 40 nearest neighbours, and the 13 principal components $^{37,38}\!;\,11$ clusters were detected. We then compared the composition of each cluster in terms of total number and percentage of wild-type and IL-27RA-knockout cells and found cluster 5 to be enriched for wild-type CD8 TILs cells (P = 0.0357, one sample t-test, Extended Data Fig. 3c, d). Projecting the IL-27 co-inhibitory gene module onto the scRNA-seq data highlighted clusters 4 and 5 (CD4 and CD8, respectively) (Extended Data Fig. 3e), further showing that in addition to the decrease in the expression of the co-inhibitory receptors: PD-1, TIM-3, LAG-3 and TIGIT (Fig. 1e), a significant decrease in the total IL-27 co-inhibitory gene module signature score is observed with lack of IL-27 signalling (P = 0.01, t-test, Extended Data Fig. 3f). Last, we searched for differentially expressed genes between clusters 4 and 5 and the rest of the clusters using a nonparametric binomial test³⁵.

Signature analysis of other states of T cell non-responsiveness. Given that orthogonal approaches were used to generate the various signatures, we first addressed the robustness of each signature before the comparative analysis. First, to address some of the concerns regarding the definition of these signatures, we sub-sampled the genes in each of the signatures and observed the resulting changes by projection on the single-cell data. These changes were quantified by randomly selecting decreasing subsets of genes from each signature (100%, 90%...30%) and calculating the average silhouette width of the cells that scored high for the different generated signatures, based on Euclidian distance between the principal component values used to generate the *t*-SNE plot. This analysis shows that the signatures are relatively resilient to this procedure up to 60% of the original signature (Extended Data Fig. 4e).

Second, we calculated a signature P value per cell. The P value is calculated by generating random sets of signatures that are composed of genes with a similar average and variance expression levels as the original signature. This was followed by comparing the generated scores to the score obtained from the original signature. Cells that had a statistically significant score (adjusted P < 0.05) were marked by a plus symbol '+' (Extended Data Fig. 4f).

For viral exhaustion, a microarray dataset 10 was downloaded, followed by RMA. A signature of viral exhaustion was defined as the genes that are differentially expressed between chronic and acute viral infection on day 15 and day 30. Genes were ranked based on a t-test statistic and fold change, each gene rank was then adjusted for multiple hypotheses testing using FDR. A threshold of fold change > 1.1 and FDR < 0.2 was applied.

For antigen-specific tolerance, data 11 were downloaded. Two groups were defined, group 1 that includes the PBS and 0.008 μg treated samples (treatment number 1) versus group 2–80 μg (treatment number 5 and 6). After \log_2 transformation and quantile normalization, the Limma package was used to estimate the fold changes and standard errors by fitting a linear model for each gene for the assessment of differential expression. Genes with $P\!<\!0.05$ were selected: 1,845 genes were upregulated of which 88 were defined as cytokine and cell surface molecules 26,39,40 .

For antigen non-specific tolerance, data¹² were downloaded. Robust multi-array average (RMA) and quantile normalization were applied for background correction and normalization using the ExpressionFileCreator module of GenePatterns.

Differentially expressed genes were defined using signal-to-noise ratio, following FDR correction. Differentially expressed genes were identified as genes having a FDR $<\!0.2$ between mRNA expression profiles of naive CD4+ or CD4+GFP/IL-10+T cells isolated from the spleen or central lymph nodes of B6NODF1 $^{\rm IL-10-GFP}$ mice following nasal treatment with anti-CD3, which attenuates the of progressive phase of experimental autoimmune encephalomyelitis.

For cancer, data³ were obtained. In brief, mRNA samples from CD8+TIM-3-PD-1- (double negative) TILs, CD8+TIM-3-PD-1+(single positive), and CD8+TIM-3+PD-1+ (double positive) TILs were measured using Affimetrix GeneChip Mouse Genome 430 2.0 Arrays, expression values were RMA normalized, corrected for batch effects using ComBat³6 and gene-specific intensities were then computed by using the maximal prob intensity per gene, values were transferred to log-space by taking $\log_2(\text{intensity})$. Differentially expressed genes were defined as genes with either an FDR-corrected t-test P value smaller or equal to 0.2 computed between the double negative and double positive subpopulations and a fold-change of at least 1.5 between the two subpopulations.

The IL-27 co-inhibitory gene module was defined as a union of the overlap between the IL-27-driven gene program (1,201 genes; see 'Microarray processing and analysis' section) and each of the four different states of T cell non-responsiveness mentioned above (272 genes, Supplementary Table 2).

For the IL-27RA-knockout signature, mRNA samples from FACS sorted CD8⁺ TILs from wild-type and IL-27RA-knockout mice bearing B16 melanomas were measured an adaptation of the SMART-Seq 2 protocol²⁸ (see 'RNA expression profiling of tumour infiltrating cells' section). Differentially expressed genes were defined as genes with either an FDR-corrected *t*-test *P* value smaller or equal to 0.2 computed between the wild-type and IL-27RA-knockout, or a fold-change of at least 1.5 between the two subpopulations. IL-27RA-knockout signature was defined as 929 differentially expressed genes in IL-27RA-knockout CD8⁺ TILs compared to wild-type CD8⁺ TILs.

Single-cell gene signature computation. As an initial step, the data were scaled (*z*-score across each gene) to remove bias towards highly expressed genes. Given a gene signature (list of genes), a cell-specific signature score was computed by first sorting the normalized scaled gene expression values for each cell followed by summing up the indices (ranks) of the signature genes. For gene-signatures consisting of an upregulated and downregulated set of genes, two ranking scores were obtained separately, and the downregulated associated signature score was subtracted from the upregulated generated signature score. A contour plot was added on top of the *t*-SNE space, which takes into account only those cells that have a signature score above the mean to further emphasis the region of highly scored cells.

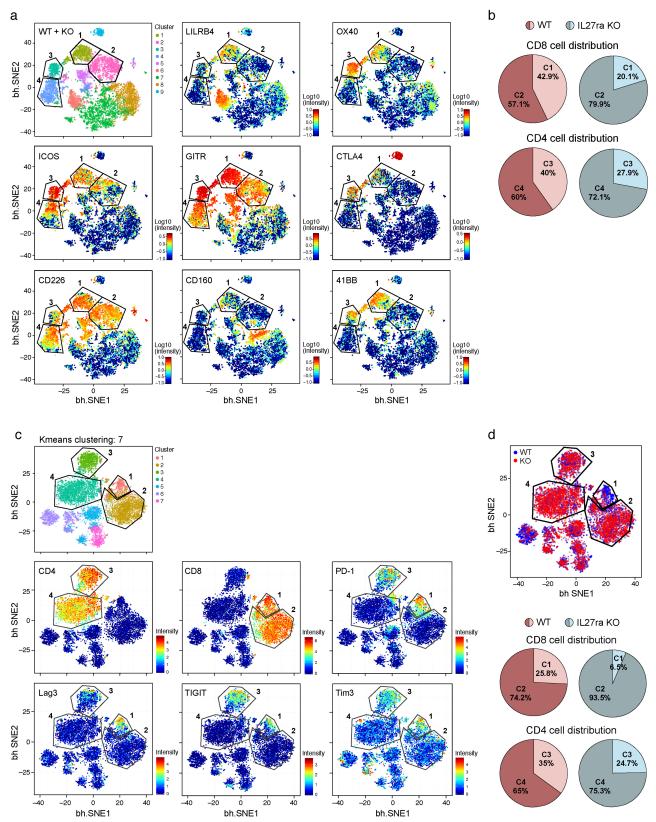
Network construction. Networks were generated using Cytoscape version 3.2.1⁴¹. The network model is based on coupling in vitro RNA-seq gene expression data of naive CD8⁺ T cells from knockout (PRDM1 or c-MAF) and wild-type controls stimulated in the presence of IL-27 and previously published ChIP-seq data for c-MAF and predicted PRDM1-binding sites by motif scan. More specifically, differentially expressed genes between wild-type control and knockout were defined using the function that fits multiple linear models from the Bioconductor package limma in R²⁶ with FDR <0.05. We used published c-MAF ChIP-seq data¹⁹ and PRDM1 ChIP-seq data¹⁶. In addition, potential PRDM1-binding sites were detected using FIMO (MEME suite; http://meme-suite.org/doc/fimo. html). Association to gene promoters was based on the following thresholds (upstream = 5,000, downstream = 500 of transcription start site) and the overlap with the co-inhibitory module was found to be significant (P = 0.009 hypergeometric, background of 20,000 genes). In the network presentation, we visualize all the genes that are part of the IL-27 inhibitory module (Fig. 4a, Extended Data Fig. 6e). **Reporting summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Sequence data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) with the accession code GSE113968. All other data are available from the corresponding authors upon reasonable request.

- 25. Wende, H. et al. The transcription factor c-Maf controls touch receptor development and function. *Science* **335**, 1373–1376 (2012).
- Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 3, Article3 (2004).
- 27. Reich, M. et al. GenePattern 2.0. Nat. Genet. 38, 500–501 (2006).
- Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352, 189–196 (2016).
- Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111 (2009).
- Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011).
- Anders, S. & Huber, W. Differential expression analysis for sequence count data. Genome Biol. 11, R106 (2010).



- Shalek, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498, 236–240 (2013).
 Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protocols* 9, 171–181 (2014).
- 34. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memoryefficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).
- 35. Shekhar, K. et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell 166, 1308-1323 (2016).
- 36. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118-127
- 37. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech. https://doi.org/10.1088/1742-5468/2008/10/P10008 (2008).
- Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitorlike cells that correlate with prognosis. Cell 162, 184-197 (2015).
- Smyth, G. K. in Bioinformatics and Computational Biology Solutions using R and Bioconductor. Statistics for Biology and Health (eds. Gentleman, R. et al.) 397–420 (Springer, New York, 2005). 40. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression
- Omnibus (GEO) and BioConductor. Bioinformatics 23, 1846-1847 (2007).
- 41. Lopes, C. T. et al. Cytoscape Web: an interactive web-based network browser. Bioinformatics 26, 2347-2348 (2010).

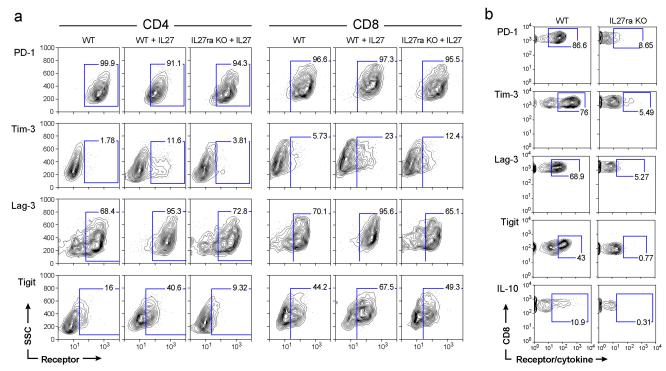


Extended Data Fig. 1 | See next page for caption.



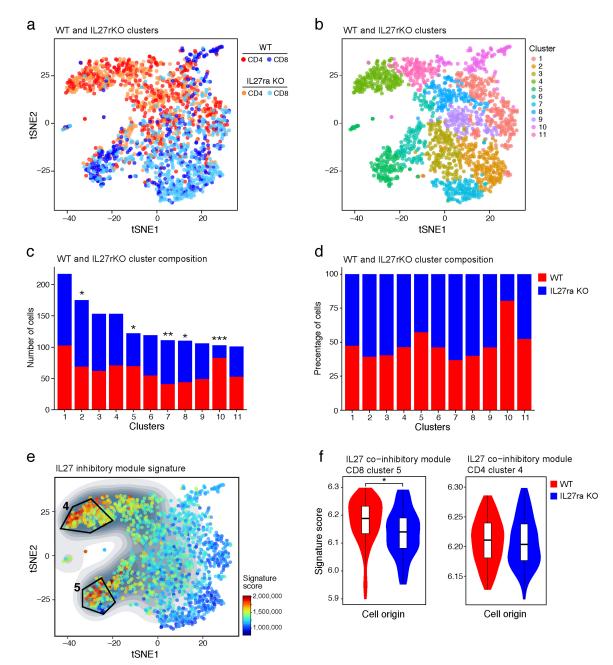
Extended Data Fig. 1 | CyTOF analysis of co-inhibitory and costimulatory receptor co-expression in TILs. a, TILs were collected from B16F10 melanoma tumour-bearing wild-type and IL-27RA-knockout mice from Fig. 1b and analysed using CyTOF (5,000 cells from each). CyTOF data were analysed using viSNE. Applying k-means clustering with k=9 on the CyTOF data resulted in a clear distinction between clusters 1, 2, 3 and 4. Polygons indicating clusters 1 and 2 (in CD8+ T cells), and 3 and 4 (in CD4+ T cells) are shown. Individual panels show expression of the indicated markers. b, Pie charts show the distribution of wild-type or IL-27RA-knockout CD8+ and CD4+ TILs in clusters 1 and 2 (C1 and C2)

of CD8⁺ TILs and clusters 3 and 4 (C3 and C4) of CD4⁺ TILs as defined in Fig. 1d. c, Independent data of wild-type and IL-27RA-knockout TILs samples from that shown in Fig. 1 (5,000 cells from each). Applying k-means clustering with k=7 on the CyTOF data resulted in a clear distinction between clusters 1, 2, 3 and 4. Polygons indicating clusters 1 and 2 (in CD8⁺ T cells), and 3 and 4 (in CD4⁺ T cells) are shown. d, viSNE plot highlighting the distribution of cells from wild-type (blue) and IL-27RA-knockout (red) mice in CD8⁺ TILs clusters 1 and 2 and CD4⁺ TILs clusters 3 and 4. Pie charts show the distribution of wild-type or IL-27RA-knockout CD8⁺ and CD4⁺ TILs in each cluster.



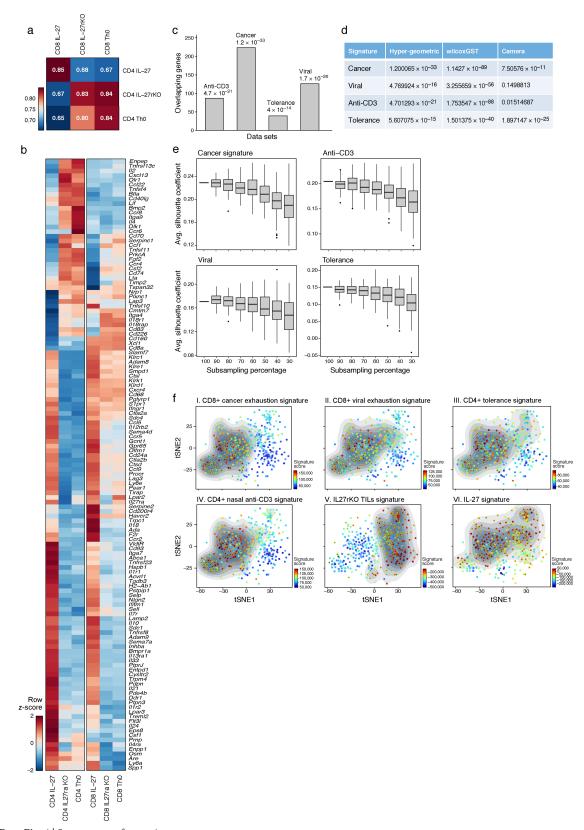
Extended Data Fig. 2 | IL-27 induces multiple co-inhibitory receptors on CD4 $^+$ and CD8 $^+$ T cells. a, Naive T cells from wild-type or IL-27RA-knockout mice were stimulated in vitro with anti-CD3/CD28 in the presence or absence of IL-27. Expression of co-inhibitory receptors was determined by flow cytometry. Representative data of three biologically independent experiments are shown. b, Expression of PD-1, TIM-3,

LAG-3, TIGIT and IL-10 on CD8⁺ TILs obtained from wild-type and IL-27RA-knockout mice bearing B16F10 melanoma was determined by flow cytometry. Thy1.1-IL-10 reporter mice crossed with wild-type and IL-27RA-knockout mice were used for IL-10 expression analysis. Representative data of three biologically independent experiments are shown.



Extended Data Fig. 3 | scRNA-seq expression analysis of wild-type and IL-27RA-knockout TILs. a, TILs were obtained from B16F10 melanoma tumour-bearing wild-type (707 and 825 for CD4⁺ and CD8⁺, respectively) and IL-27RA-knockout (376 and 394 for CD4⁺ and CD8⁺, respectively) mice as in Fig. 1e. t-SNE plot shows the presence of wild-type and IL-27RA-knockout CD4⁺ and CD8⁺ TILs as indicated. b, Clustering using the Louvain–Jaccard method (40 nearest neighbours and 13 principal components³⁷). c, The composition of each cluster in terms of total number (c) and percentage (d) of wild-type (red) and IL-27RA-knockout (blue) cells. *P < 0.05, **P < 0.01, ***P < 0.001, one sample t-test.

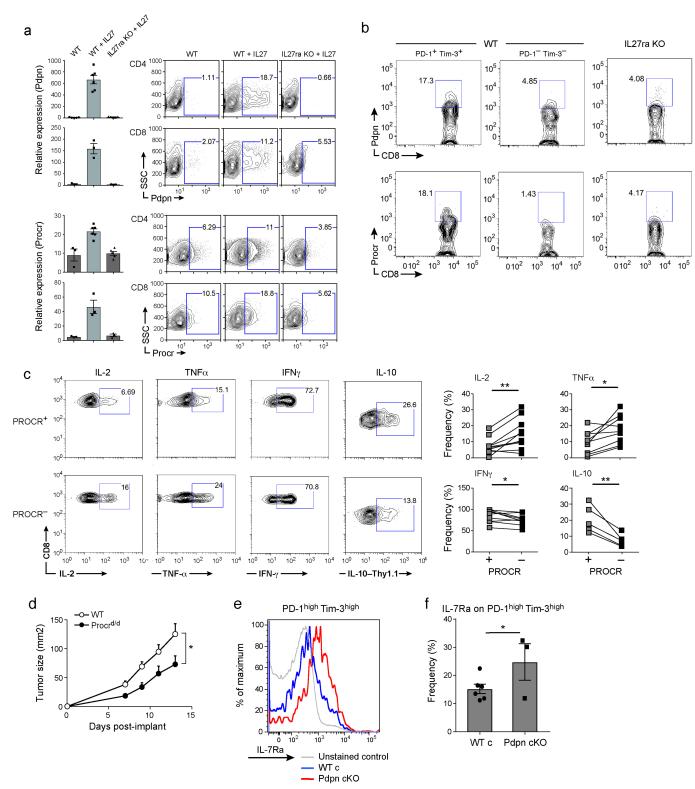
e, Projection of the IL-27 co-inhibitory module signature on the scRNA-seq data. The contour plot marks the region of highly expressing cells by taking into account only those cells that have an expression value above the mean. **f**, Violin and box plots displaying the distribution of the IL-27 co-inhibitory module signature score compared between wild-type (72 and 98 for CD4⁺ and CD8⁺, respectively) and IL-27RA-knockout (85 and 77 for CD4⁺ and CD8⁺, respectively) cells in clusters 4 and 5 (CD4⁺ and CD8⁺, respectively). *P = 0.01, one-sided t-test. The top and bottom hinges in the boxplot correspond to the first and third quartiles, and the horizontal line denotes the median.



Extended Data Fig. 4 \mid See next page for caption.

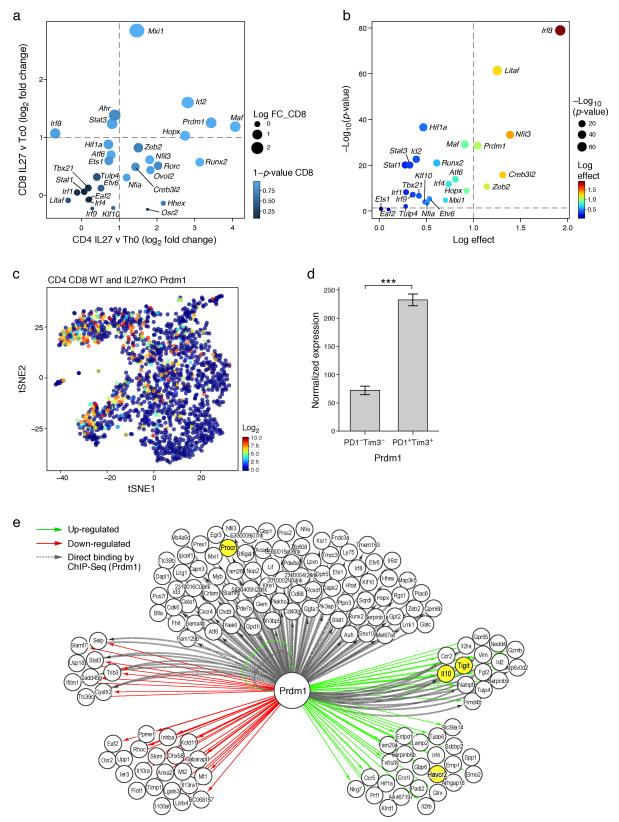
Extended Data Fig. 4 | Overlap of the IL-27-induced gene program with signatures from four states of T cell impairment, tolerance and **dysfunction.** a, Pearson correlation between wild-type CD4⁺ and CD8⁺ T cells for the 1,201 genes that were differentially expressed between wild-type CD4⁺ T cells stimulated in the presence or absence of IL-27 (FDR < 0.05). **b**, Expression profile of 118 differentially expressed genes (from a) encoding cell-surface receptors and cytokines are shown as a heat map. c, The IL-27-induced gene program (1,201 genes) was compared to T cell signatures obtained from four states of T cell non-responsiveness. Number of overlapping genes between the IL-27 gene program and each signature is depicted. ***P < 0.001, hypergeometric test: nasal anti-CD3 4.7×10^{-21} ; cancer 1.2×10^{-33} ; antigen-specific tolerance 4×10^{-14} ; and viral exhaustion 1.7×10^{-26} . **d**, P-value statistics for the significance of the overlap between the IL-27-induced gene program (1,201) and genes induced in other states of T cell non-responsiveness using WilcoxGST and camera. e, Gene signatures from c were sub-sampled and projected onto the CD8⁺ single-cell TIL data. Changes were quantified by randomly

selecting decreasing subsets of genes from each signature and calculating the average silhouette width of cells that scored high for the different generated signatures based on Euclidian distance between the principal component values used to generate the *t*-SNE plot. The top and bottom hinges in the boxplot correspond to the first and third quartiles, and the horizontal line denotes the median (Methods). f, Panels I–V, t-SNE plots of the 588 $\mathrm{CD8^{+}}$ single-cell TILs (dots) obtained from wild-type mice bearing B16F10 melanoma tumour. Cells are coloured by their signature score. The score reflects the relative average expression of the genes in the overlap of the IL-27 gene signature with the signatures for each of the indicated states of T cell non-responsiveness. Panel VI is a projection of a signature of the differentially expressed genes between CD8⁺ TILs from wild-type and IL-27RA-knockout mice bearing B16F10 melanomas (Methods). The contour plot marks the region of highly scored cells by taking into account only those cells that have a signature score above the mean score. Cells that had a statistically significant score (adjusted P < 0.05) are marked by a plus symbol (Methods).



Extended Data Fig. 5 | Characterization of the role of PDPN and PROCR in CD8+ TILs. a, PDPN and PROCR protein and mRNA expression was determined in T cells from wild-type and IL-27RA-knockout mice stimulated with anti-CD3/CD28 in the presence or absence of IL-27. CD4+ cells were analysed at 96 h and CD8+ cells at 72 h. Data are mean \pm s.e.m. from representative flow cytometry and qPCR data from biologically independent animals. b, Representative flow cytometry data of three independent experiments showing PDPN and PROCR expression in PD-1+TIM-3+CD8+ and PD-1-TIM-3-CD8+ TILs obtained from wild-type and IL-27RA-knockout mice bearing B16F10 melanoma. c, TILs from wild-type mice bearing B16F10 melanoma were stimulated with phorbol myristate acetate and ionomycin. Cytokine production in PROCR+ or

PROCR $^-$ CD8 $^+$ TILs is shown. Thy1.1-IL-10 reporter mice were used for IL-10 expression analysis. Data are mean \pm s.e.m. from biologically independent animals. $^*P < 0.05, \, ^{**}P < 0.01,$ paired t-test. $\mathbf{d},$ CD8 $^+$ T cells (5×10^5) from wild-type or $Procr^{d/d}$ mice were transferred along with 1×10^6 wild-type CD4 $^+$ T cells to RAG1-knockout mice (n=5). On day 2, 5×10^5 B16F10 cells were implanted. Data are mean \pm s.e.m. $^*P < 0.05,$ repeated measures ANOVA, Sidak's multiple comparisons test. $\mathbf{e},$ TILs were obtained from wild-type and PDPN cKO mice bearing B16F10 melanoma and stained for the expression of IL-7RA. Representative flow cytometry data from three independent animals. $\mathbf{f},$ Summary data of IL-7RA expression are from biologically independent animals. Data are mean \pm s.e.m. $^*P < 0.05,$ one-sided t-test.

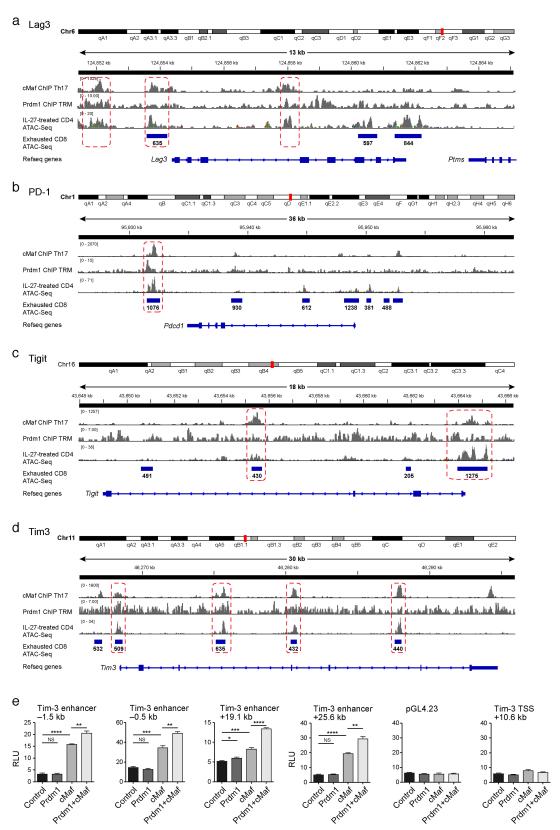


Extended Data Fig. 6 | See next page for caption.



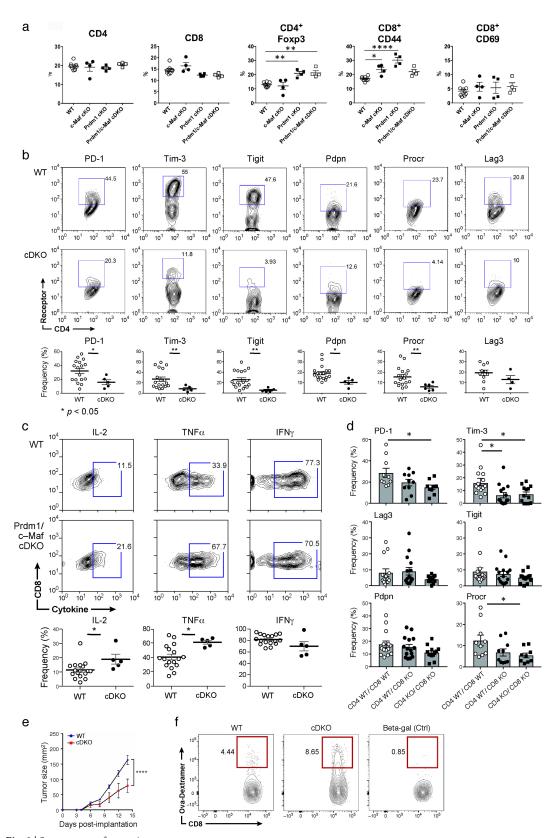
Extended Data Fig. 6 | PRDM1 is a candidate regulator of the coinhibitory module. a, The \log_2 fold change in RNA levels between naive CD4+ or CD8+ T cells simulated with or without IL-27. Data are from two independent experiments. Transcription factors that are part of the IL-27 co-inhibitory module are shown (differentially expressed transcription factors were annotated as genes with FDR-corrected ANOVA < 0.05). b, Transcription factors that are both in the IL-27 co-inhibitory module and are also overexpressed in clusters 4 and 5 in the single-cell data (clusters that were enriched for the IL-27 signature; Extended Data Fig. 3e, f). Differentially expressed genes between clusters 4 and 5 and the rest of the clusters were determined using binomcount.test (binomial distribution, Methods). The log effect corresponds to log proportion of expressing cells and the *P* value is calculated by the probability of finding

n or more cells positive for the gene in clusters 4 and 5 given the fraction in the rest of the clusters. **c**, t-SNE plot of Fig. 1e showing the expression of PRDM1 in wild-type (707 and 825 for CD4⁺ and CD8⁺, respectively) and IL-27RA-knockout (376 and 394 for CD4⁺ and CD8⁺, respectively) cells. **d**, Normalized RNA expression levels of PRDM1 in PD-1⁻TIM-3⁻ (n = 3) and PD-1⁺TIM-3⁺ (n = 3) CD8⁺ TILs. Data are mean \pm s.e.m. ***P = 0.0004, two-sided t-test. **e**, Network model based on RNA-seq gene expression data of naive CD8⁺ T cells from $Prdm1^{fl/fl}$ (WT) or $Cd4^{cre}Prdm1^{fl/fl}$ (PRDM1 cKO) mice stimulated in the presence of IL-27 and actual binding events (ChIP-seq data for PRDM1)¹⁹. Green arrows designate genes upregulated by PRDM1, red arrows designate genes downregulated by PRDM1, and dashed grey arrows denote binding events.



Extended Data Fig. 7 | Genomic tracks surrounding the co-inhibitory molecules. a–d, LAG-3 (a), PD-1 (b), TIGIT (c) and TIM-3 (d) with overlay of ChIP–seq data of PRDM1 16 and c-MAF 19 and ATAC-seq data of naive CD4 $^+$ cells induced with IL-27 for 72 h and ATAC-seq data of CD8 $^+$ T cells 27 days after chronic viral infection 22 . Regions of binding sites common to both PRDM1 and c-MAF are indicated by the dotted rectangles. e, Luciferase activity in 293T cells transfected with pGL4.23

luciferase reporters for depicted enhancers of TIM-3 together with empty vector (control), constructs encoding PRDM1, c-MAF or both. Firefly luciferase activity was measured 48 h after transfection and is presented relative to constitutive *Renilla* luciferase activity. Data are mean \pm s.e.m. from biologically independent experiments. *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.001, one-way ANOVA and Tukey's multiple comparisons test.

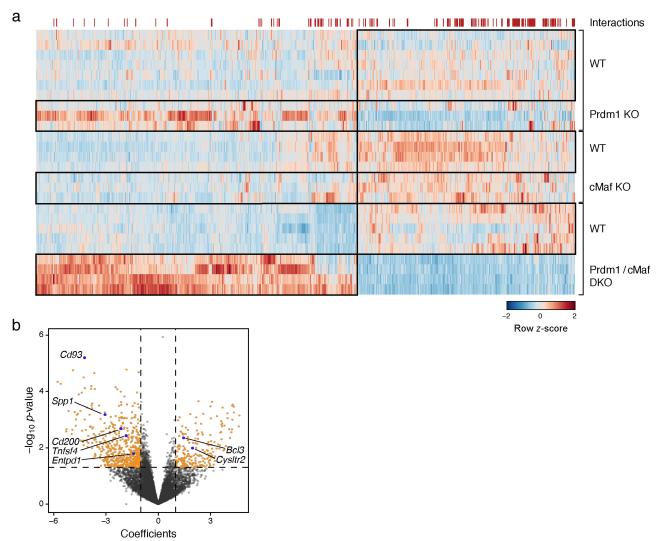


Extended Data Fig. 8 \mid See next page for caption.



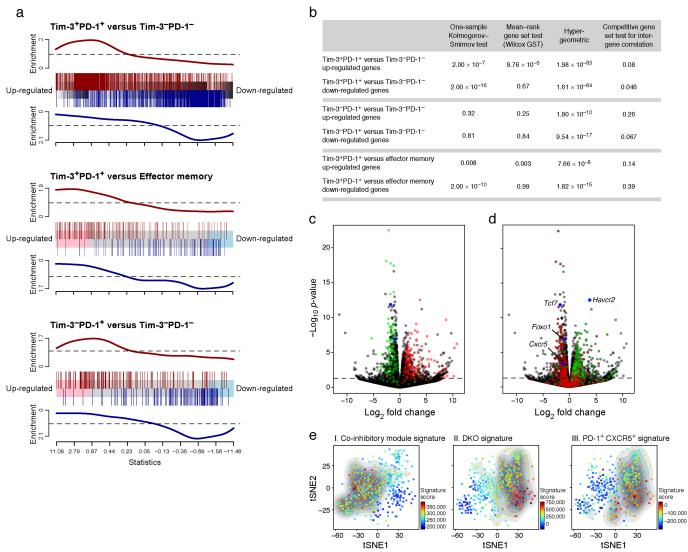
Extended Data Fig. 8 | Immune characterization of PRDM1 cKO, c-MAF cKO and PRDM1/c-MAF cDKO before and after tumour challenge. a, Analysis of steady-state immune system in wild-type, c-MAF cKO, PRDM1 cKO and PRDM1/c-MAF cDKO. Summary data for CD4, CD8, FOXP3, CD44, CD62L and CD69 expression in spleen from wild-type, c-MAF cKO, PRDM1 cKO and PRDM1/c-MAF cDKO mice. Data are mean \pm s.e.m. from biologically independent animals. *P < 0.05, **P < 0.01, ****P < 0.0001, one-way ANOVA and Tukey's multiple comparisons test. b, Co-inhibitory receptor expression in CD4⁺ TILs from PRDM1/c-MAF cDKO mice Top, representative flow cytometry data from three independent experiments for TILs from wild-type and PRDM1/c-MAF cDKO stained for PD-1, TIM-3, TIGIT, PDPN and PROCR expression. Bottom, summary data. Data are mean \pm s.e.m. from biologically independent animals. *P < 0.05, two-sided t-test. c, Top, representative flow cytometry data from three independent experiments

showing cytokine production from CD8⁺ TILs from wild-type and cDKO bearing B16F10 melanoma. Bottom, summary data. Data are mean \pm s.e.m. from biologically independent animals. *P < 0.05, two-sided t-test. **d**, Co-inhibitory receptor expression on CD8⁺ TILs sorted from B16-OVA-bearing RAG1-knockout mice that were transferred with PRDM1/c-MAF cDKO (n=4) or wild-type (n=4) CD4⁺ and CD8⁺ T cells as indicated. Data are mean \pm s.e.m. from biologically independent animals. *P < 0.05, one way ANOVA and Tukey's multiple comparisons test. **e**, RAG1-knockout mice were transferred with either wild-type or cDKO CD4⁺ and CD8⁺ (2:1 CD4:CD8 ratio), followed by subcutaneous injection of MC38-OVA. Data are mean \pm s.e.m. ****P < 0.0001, repeated measures ANOVA, Sidak's multiple comparisons test. On day 14 after tumour implantation, mice were euthanized and TILs, spleen and draining lymph nodes were obtained. **f**, The frequency of antigen-specific CD8⁺ T cells in the draining lymph nodes of mice in **e**.



Extended Data Fig. 9 | Examination of additive and non-additive (synergistic) effects of PRDM1 and c-MAF. a, A heat map showing all 940 differentially expressed genes between wild-type (n=5) and cDKO (PRDM1/c-MAF, n=4) mice, and their expression in single knockout (PRDM1 control n=7, PRDM1 knockout n=3, c-MAF control n=4 and c-MAF-knockout n=3) mice. The red markings at the top indicate genes on expression of which the two knockouts have a statistically significant

 $(P\,{<}\,0.05)$ non-additive effect in the cDKO (149 out of 940 differentially expressed genes). b, Volcano plot of the analysis in a for global gene expression. Genes whose expression in the two single knockouts have a statistically significant ($P\,{<}0.05)$ non-additive effect in the cDKO (1,144 out of 12,906 genes) and had an absolute coefficient value ${>}$ 1 (779 out of 1,144) are shown in orange.



Extended Data Fig. 10 | Comparison of gene expression between PRDM1/c-MAF cDKO TILs and CD8+ TILs populations from wild-type mice. a, Barcode enrichment plot displaying two gene sets in a ranked gene list. The ranked gene list was defined as fold change in gene expression between PRDM1/c-MAF cDKO and wild-type CD8+ TILs. The three gene sets consist of differentially expressed genes between: PD-1+TIM-3+CD8+ (n=3) and PD-1-TIM-3-CD8+ (n=3) TILs, PD-1+TIM-3+CD8+ (n=3) TILs and memory CD8+ (n=3), and PD-1+TIM-3-CD8+ (n=3) and PD-1-TIM-3-CD8+ TILs. **b**, This analysis was followed by four statistical tests (one-sample Kolmogorov-Smirnov test, mean-rank gene set test (WilcoxGST), hypergeometric, and competitive gene set test accounting for inter-gene correlation) for enrichment of these signatures in the cDKO expression profile.

c, Wild-type versus cDKO volcano plot. Green indicates genes that were upregulated in the PD-1⁻TIM-3⁻CD8⁺ (double negative) TILs and red indicates genes that were upregulated in the PD-1⁺TIM3⁺CD8⁺ (double positive) TILs. d, Wild-type versus cDKO volcano plot. Red indicates genes that were upregulated in PD-1⁺CXCR5⁺CD8⁺ T cells and green indicates genes that were upregulated in PD-1⁺CXCR5⁻CD8⁺ T cells in chronic LCMV infection²³. e, A *t*-SNE plot of the 588 CD8⁺ TILs obtained from wild-type mice bearing B16F10 melanoma tumours, coloured by the relative signature score for the co-inhibitory module (272 genes, Supplementary Table 2), the cDKO signature (shown in g), and the PD-1⁺CXCR5⁺CD8⁺ T cell signature from chronic virus infection²³. The contour plot marks the region of highly scored cells by taking into account only those cells that have a signature score above the mean.

natureresearch

Corresponding author(s):	Vijay K.Kuchroo	
☐ Initial submission ☐	Revised version	Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1	_	
1.	Samn	CIZO

Describe how sample size was determined.

At least 5 animals of target gene knock out and control mice were used to adequately power biological validation experiments throughout the article. Statistical differences provide the rationale for sufficiency of the sample sizes.

2. Data exclusions

Describe any data exclusions.

None

3. Replication

Describe whether the experimental findings were reliably reproduced.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No randomization was used as animals were genotyped prior to use.

confirmed findings were reliably reproduced.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Since mice were genotyped before the experiments , no randomization was used. However, investigators injected tumor randomly and tumor-size was assessed randomly to avoid any bias as much as possible. We will indicate this in the method section in the revised version.

Experiments were repeated multiple times to ensure reproducibility of results and

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

,	C C 1
า/ล	l Confirmed

\boxtimes	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
\times	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

Ш	\times	1	A statement indicating how ma	any times (each experiment w	vas replicated

1	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more
	complex techniques should be described in the Methods section)

$ \rangle$	/	Δο	loccri	ntion	of ar	ny acciim	nntions c	or co	orrections,	such	ac an	adi	ilistment :	for mi	ıltin	le com	narice	٦n
11/	\sim	\neg	IC3CI I	puon	Oi ai	iy assuii	iptions t)	ni celions,	Jucii	as an	au	justilielit	101 1110	aιτιρ	ic com	parise	711.

The test results (e.g. <i>P</i> values) given as exact values whenever possible and with confidence interva	intervals	confidence i	le and with	possible	values whenever	given as exact	P values)	e.g.	results	The test	\mathbb{N}
---	-----------	--------------	-------------	----------	-----------------	----------------	-----------	------	---------	----------	--------------

🛮 🔀 A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)

Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.



A Cdk9-PP1 switch regulates the elongationtermination transition of RNA polymerase II

Pabitra K. Parua¹, Gregory T. Booth², Miriam Sansó^{1,4}, Bradley Benjamin¹, Jason C. Tanny³, John T. Lis² & Robert P. Fisher^{1*}

The end of the RNA polymerase II (Pol II) transcription cycle is strictly regulated to prevent interference between neighbouring genes and to safeguard transcriptome integrity¹. The accumulation of Pol II downstream of the cleavage and polyadenylation signal can facilitate the recruitment of factors involved in mRNA 3'-end formation and termination², but how this sequence is initiated remains unclear. In a chemical-genetic screen, human protein phosphatase 1 (PP1) isoforms were identified as substrates of positive transcription elongation factor b (P-TEFb), also known as the cyclin-dependent kinase 9 (Cdk9)-cyclin T1 (CycT1) complex³. Here we show that Cdk9 and PP1 govern phosphorylation of the conserved elongation factor Spt5 in the fission yeast Schizosaccharomyces pombe. Cdk9 phosphorylates both Spt5 and a negative regulatory site on the PP1 isoform Dis2⁴. Sites targeted by Cdk9 in the Spt5 carboxy-terminal domain can be dephosphorylated by Dis2 in vitro, and dis2 mutations retard Spt5 dephosphorylation after inhibition of Cdk9 in vivo. Chromatin immunoprecipitation and sequencing analysis indicates that Spt5 is dephosphorylated as transcription complexes traverse the cleavage and polyadenylation signal, concomitant with the accumulation of Pol II phosphorylated at residue Ser2 of the carboxy-terminal domain consensus heptad repeat⁵. A conditionally lethal Dis2-inactivating mutation attenuates the drop in Spt5 phosphorylation on chromatin, promotes transcription beyond the normal termination zone (as detected by precision run-on transcription and sequencing⁶) and is genetically suppressed by the ablation of Cdk9 target sites in Spt5. These results suggest that the transition of Pol II from elongation to termination coincides with a Dis2-dependent reversal of Cdk9 signalling—a switch that is analogous to a Cdk1-PP1 circuit that controls mitotic progression⁴.

In metazoans and fission yeast, the inhibitory phosphorylation of PP1 by Cdk1 and its abrupt reversal promote orderly progression through mitosis⁴. In human extracts, analogue-sensitive Cdk9 modified PP1 β and PP1 γ on conserved, carboxy-terminal sites analogous to the PP1 α residue labelled by Cdk1^{3,7}. Of the two fission-yeast PP1 isoforms, Dis2 and Sds21, only Dis2 has the potential for inhibition by cyclin-dependent kinases (CDKs), through phosphorylation of Thr316; the sole budding-yeast PP1 catalytic subunit, Glc7, lacks this regulatory site^{8,9} (Fig. 1a). Purified S. pombe Cdk9 phosphorylated Dis2 in vitro but not Sds21 (Fig. 1b); labelling was diminished by a T316A substitution, but not by a Dis2-inactivating point mutation 10. Dis2 is regulated specifically by Cdk9 in vivo. Treatment of analogue-sensitive cdk9as cells, but not wild-type cells or cells with analogue-sensitive versions of the transcriptional CDKs Mcs6 (an orthologue of metazoan Cdk7) or Lsk1 (orthologue of Cdk12), with the bulky adenine analogue 3-MB-PP1 (which inhibits all three analogue-sensitive CDKs¹¹), decreased Thr316 phosphorylation of chromatin-associated Dis2 (Fig. 1c, Extended Data Fig. 1a).

The previously known target of S. pombe Cdk9 is Thr1 in the nonapeptide carboxy-terminal domain (CTD) repeat $T_1P_2A_3W_4N_5S_6G_7S_8K_9$ of $\mbox{Spt} S^{12,13}$. A phosphopeptide with this sequence was dephosphorylated

in vitro by PP1 purified from bacteria (Extended Data Fig. 1b) or immunoprecipitated from yeast (Extended Data Fig. 1c-e). We recovered a similar amount of Dis2 from dis2⁺ and dis2-11 cold-sensitive mutant cells, but detected activity only with the former, consistent with a previous observation that the enzyme encoded by dis2-11 was severely impaired even at a permissive temperature 10. Dis2 recovered from $sds21\Delta$ cells also had reduced activity, perhaps suggesting a contribution by Sds21 to Dis2 activation. Cdk9 treatment of wild-type Dis2 increased its reactivity with anti-Dis2-pT316 antibodies and diminished its phosphatase activity, whereas the same treatment had little effect on the constitutively higher activity of Dis2(T316A) or the lower activity of Dis2(T316D) (Fig. 1d, Extended Data Fig. 1d). Together, the results indicate that Dis2 is a target of negative regulation by Cdk9 and a potential Spt5 phosphatase, although it could also promote dephosphorylation of transcriptional CDK substrates by activating protein phosphatase 2A (PP2A), as it does in mitosis⁴. Either arrangement predicts switch-like behaviour of pSpt5 turnover and Pol II dynamics in response to changes in Cdk9 activity (Fig. 1e).

Consistent with this prediction, pSpt5, measured with a phosphospecific antibody against the *S. pombe* Spt5-derived phosphopeptide described above 13 , was rapidly lost after addition of 3-MB-PP1 to $cdk9^{as}$ cells (half-life ≈ 20 s) (Fig. 2a). The transcription machinery also responds rapidly to Cdk9 shutoff. Precision run-on transcription and sequencing (PRO–seq) analysis in $cdk9^{as}$ cells revealed Pol II slowing within 30 s of inhibitor addition 14 . The rate of pSpt5 decay after Cdk9 inhibition was approximately twofold slower in dis2-11 relative to $dis2^+$ cells at a permissive temperature of 30 °C, and reduced by approximately fourfold in dis2-11 cells shifted to 18 °C before drug addition (Fig. 2a, Extended Data Fig. 2a). Rapid dephosphorylation was restored by expression of active Dis2 in dis2-11 cells (Fig. 2b, Extended Data Fig. 1e). Therefore, the maximal rate of pSpt5 turnover depends on Dis2 activity in vivo.

Inhibition of Lsk1 had no effect on pSpt5 (Extended Data Fig. 2b) but diminished phosphorylation of the Pol II CTD repeat (consensus: $Y_1S_2P_3T_4S_5P_6S_7$) on Ser2 (pS2, Extended Data Fig. 2c). This mark was refractory to Lsk1 inhibition at 37 °C in cells with a temperature-sensitive mutation in $fcp1^{15}$, which encodes a pS2-specific phosphatase^{16,17}. The rate of pS2 decay was unaffected by dis2-11 (Extended Data Fig. 2d) and, conversely, Fcp1 inactivation had no effect on pSpt5 stability in $cdk9^{as}$ strains (Extended Data Fig. 2e). Similarly, pSpt5 turn-over was impervious to genetic inactivation of Ssu72, a Pol II CTD Ser5 phosphatase¹⁸ (Extended Data Fig. 2f). We surmise that orthogonal CDK–phosphatase pairs govern the Pol II and Spt5 phosphorylations that are implicated in elongation.

Dis2 also influences pSpt5 turnover on chromatin; in an analysis using chromatin immunoprecipitation—quantitative PCR (ChIP–qPCR), pSpt5 occupancy was nearly abolished after Cdk9 inhibition for 2 min in *dis2*⁺ cells (Extended Data Fig. 3a), but was more resistant to 3-MB-PP1 treatment in *cdk9*^{as} *dis2-11* cells at 18 °C (Fig. 2c, Extended Data Fig. 3b–d). The *dis2-11* mutant also had higher pSpt5:Spt5 ChIP signal ratios when Cdk9 remained active (Fig. 2d). Cdk9 inhibition

¹Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ²Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA.

³Department of Pharmacology and Therapeutics, McGill University, Montreal, Quebec, Canada. ⁴Present address: Cancer Genomics Group, Vall d'Hebron Institute of Oncology, Barcelona, Spain. *e-mail: robert.fisher@mssm.edu

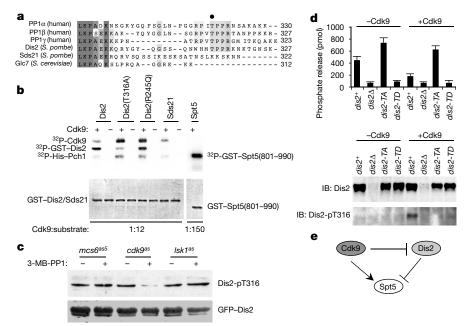


Fig. 1 | **A Cdk9–Dis2–Spt5 circuit. a**, Alignment of the C termini of human and fungal PP1 isoforms. Thr320 of PP1 α was identified as a target of Cdk1, and the analogous residues in PP1 β and PP1 γ as targets of Cdk9. This site (indicated by a dot) is conserved in fission-yeast Dis2 but not in Sds21, or in the budding yeast PP1 catalytic subunit Glc7. **b**, Phosphorylation of Dis2 Thr316 by Cdk9 in vitro. Purified, insect-cell-derived complexes of Cdk9 with its cyclin partner Pch1 were incubated at indicated molar ratios with purified, bacterially expressed GST–PP1 or GST–Spt5(801–990) (containing the CTD) after activation by CDK-activating kinase Csk1 (incubated alone in indicated lanes). In addition to wild-type Dis2, we tested Dis2(T316A) and the Dis2(R245Q) variant encoded by *dis2-11*. Autophosphorylation occurs on both Cdk9 and Pch1. Top, autoradiogram; bottom, Coomassie-stained gel to confirm equal loading. **c**, Cdk9 phosphorylates chromatin-bound Dis2 on Thr316 in vivo. Cells of the indicated strains ($mcs6^{as5}$, $cdk9^{as}$ or $lsk1^{as}$, with green

fluorescent protein (GFP)-tagged Dis2 expressed from the chromosomal $dis2^+$ locus) were treated for 10 min with 10 μ M 3-MB-PP1 or mock-treated with DMSO at 30 °C, as indicated. Chromatin extracts were immunoprecipitated with anti-GFP antibodies and probed with antibodies specific for Dis2 phosphorylated at Thr316 (Dis2-pT316) or GFP. **b, c**, Experiments were performed twice with similar results. **d,** Phosphorylation by Cdk9 decreases activity of Dis2. Top, phosphatase activity measurements on anti-Dis2 immunoprecipitates from wild-type ($dis2^+$) or mutant ($dis2\Delta$, $dis2^{T316A}$ or $dis2^{T316D}$) extracts, treated with Cdk9 or mock-treated before a phosphate release assay with Spt5-pT1 phosphopeptide. Bottom, immunoblotting (IB) was performed to verify recovery (IB: Dis2) and phosphorylation (IB: Dis2-pT316) of immunoprecipitated Dis2. Data are mean + s.d. from three biological replicates. **e**, A Cdk9-Dis2-Spt5 circuit diagram.

stimulated Dis2 recruitment to chromatin, whereas Mcs6 or Lsk1 inhibition had no consistent effect on Dis2 occupancy (Fig. 2e, Extended Data Fig. 4a–c). Dis2 ChIP signals were also enhanced by $cdk9\Delta C$, which removes a carboxy-terminal region of Cdk9 that promotes its recruitment to chromatin¹⁹, and by $cdk9^{T212A}$, which prevents Cdk9-activating phosphorylation²⁰ (Extended Data Fig. 4d). Thus Cdk9 can limit both activity and chromatin recruitment of Dis2—mutually reinforcing effects that might contribute to the switch-like behaviour of pSpt5.

Dis2 and Ssu72 are components of the cleavage and polyadenylation factor (CPF)²¹, as is Glc7, impairment of which led to termination defects in budding yeast^{22,23}. Cdk9 inhibition did not cause ectopic recruitment of the entire CPF to gene bodies; crosslinking of three other CPF subunits was not affected (Extended Data Fig. 5a). Glc7 removes Pol II CTD Tyr1 phosphorylation (pY1) to facilitate pS2dependent recruitment of termination factors²³, but pY1 ChIP signals were not increased (and pS2 was unaffected) by a dis2-11 mutation (Extended Data Fig. 5b). Recruitment of the CPF subunit Pfs2 was likewise unaffected by *dis2-11*, and thermal inactivation of Pfs2 in pfs2-11 cells did not alter pSpt5:Spt5 ratios, although it increased Pol II occupancy downstream of the cleavage and polyadenylation signal (CPS) (Extended Data Fig. 5c, d). These results support specific, direct interactions between Cdk9 and Dis2 in governing pSpt5, rather than an indirect effect of impaired 3'-end processing or loss of CPF complex integrity (while also suggesting that Glc7 in budding yeast has different regulators and targets).

In $cdk9^+$ cells with different dis2 alleles, ChIP–qPCR analysis revealed significant increases in pSpt5 (Student's t-test) in the loss-of-function mutants dis2-11, $dis2\Delta$ and $dis2^{T316D}$, relative to $dis2^+$ and

 $dis2^{T316A}$ cells (Extended Data Fig. 6a, b). The relative increases owing to dis2-11 and $dis2\Delta$ correlated with the degree of bulk pSpt5 stabilization after Cdk9 inhibition (Fig. 2a, Extended Data Fig. 6c). We suspect dis2-11 is more severely affected than $dis2\Delta$ because loss of Dis2 protein might allow more effective compensation by other phosphatases. An $sds21\Delta$ mutation did not retard pSpt5 decay, however, and Cdk9 inhibition did not increase Sds21 recruitment to chromatin (Extended Data Fig. 6d, e), indicating that Dis2 is the major PP1 isoform that regulates pSpt5 in wild-type cells, in opposition to Cdk9 and possibly in collaboration with downstream phosphatases.

Cdk9 has a rate-limiting role in Pol II elongation 14 that is probably dependent in part on Spt5, depletion of which caused Pol II accumulation in upstream gene regions in fission yeast⁵ and disrupted coupling between 3'-processing and termination in budding yeast²⁴. ChIP-seq analysis of dis2⁺ cells revealed similar distributions of total Spt5 and pSpt5 in gene bodies (Fig. 3a, b, Extended Data Fig. 7a). The patterns diverged in the region beyond the CPS, in which a prominent peak of total Spt5 corresponds to a much smaller peak of pSpt5. In *dis2-11* cells, pSpt5 signals were increased throughout gene bodies, but more so downstream of the CPS (Fig. 3a, b, Extended Data Fig 7b, c). This increased pSpt5 retention was most obvious on genes selected for high Spt5 occupancy (cluster 1 in Fig. 3a), or filtered to minimize signals from neighbouring genes (Extended Data Fig. 7c, d). A metagene analysis comparing pSpt5:Spt5 ratios between dis2+ and dis2-11 strains reveals a sharp, CPS-centred increase in the mutant (Fig. 3c, Extended Data Fig. 7d), suggesting a Dis2-dependent, switch-like transition as elongation complexes traverse this landmark. The trough of pSpt5 near the CPS in *dis2*⁺ cells coincides with a peak of Ser2-phosphorylated Pol II observed in a recent ChIP-seq analysis⁵ (Fig. 3d, Extended Data

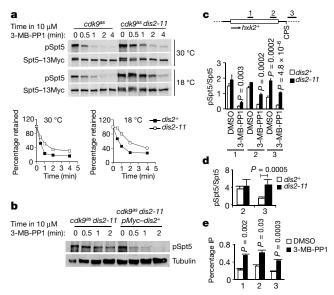


Fig. 2 | Cdk9 and Dis2 regulate Spt5 phosphorylation in vivo. a, Dis2 inactivation stabilizes pSpt5 after Cdk9 inhibition. The indicated fissionyeast strains (cdk9as spt5-13Myc or cdk9as spt5-13Myc dis2-11) were grown to mid-log phase at 30 °C and shifted to 18 °C for 10 min or not shifted before addition of 10 μ M 3-MB-PP1, after which cultures were sampled at the indicated times and subjected to immunoblot analysis with anti-pSpt5 or anti-Myc antibodies (top). Signals were quantified by fluorescence and mean values were plotted as percentage retention of pSpt5 relative to Spt5–Myc with time after addition of 3-MB-PP1 in dis2+ and dis2-11 cells (bottom, n = 2 biological replicates). **b**, Ectopic expression of wild-type Dis2 restores rapid Spt5 dephosphorylation kinetics in a *dis2* mutant. cdk9as dis2-11 cells were shifted to 18 °C and treated with 10 μM 3-MB-PP1 for the indicated times, with or without expression of Myc-Dis2 from a plasmid, before immunoblot detection of pSpt5 and tubulin (loading control). The experiment was repeated twice independently with similar results. c, Dis2 inactivation stabilizes chromatin-associated pSpt5. Either cdk9^{as} spt5-13Myc or cdk9^{as} spt5-13Myc dis2-11 cells were shifted to 18 °C and treated with 10 µM 3-MB-PP1 or mock-treated with DMSO for 2 min and subjected to ChIP-qPCR analysis at the hxk2+ locus for pSpt5 and total Spt5 (anti-Myc). The pSpt5:Spt5 ratio was plotted for dis2+ and dis2-11 cells for each treatment. d, Comparison of the pSpt5:Spt5 ratio using ChIP-qPCR analysis upstream and downstream of CPS on the $hxk2^+$ gene in $dis2^+$ and dis2-11 cells at 18 °C. **e**, Suppression of Dis2 recruitment to transcribed chromatin by Cdk9. ChIP-qPCR analysis of GFP-Dis2 crosslinking at hxk2+ in cdk9as GFP-dis2 cells treated for 10 min with 10 μM 3-MB-PP1. c-e, Data are mean + s.d. from three biological replicates; P values from Student's t-test are indicated between wild-type ($dis2^+$) and mutant (dis2-11) cells (\mathbf{c}, \mathbf{d}), or between 3-MB-PP1 and DMSO treatment (e).

Fig. 8a-c), a reciprocal relationship consistent with independent pSpt5 and pS2 regulation.

There is a transient drop in total Spt5 occupancy near the CPS (Fig. 3a, b) that is also seen in metagene plots of Pol II⁵ (Extended Data Fig. 8a) and of Spt5 in budding yeast, in which it corresponds to a peak of crosslinking of Spt5 to the nascent transcript²⁴. Although an exchange of phosphorylated Spt5 for unphosphorylated Spt5 is formally possible, active dephosphorylation near the CPS seems more likely a priori, given the similarities in Pol II and Spt5 distribution and the tight association of Spt5 with the Pol II clamp²⁵. Consistent with a phosphatase-dependent mechanism, the dis2-11 mutation caused significantly greater pSpt5 stabilization ($P = 2.8 \times 10^{-16}$, two-sided Student's *t*-test) in the 500-bp region downstream of the CPS than in the 500 bp upstream (Extended Data Fig. 8d-f). Moreover, the *dis2-11* phenotype is due in part to Spt5; replacement of Thr1 with alanine in a truncated, seven-repeat Spt5 CTD $(spt5-(T1A)_7)$, which by itself imparts sensitivity to cold²⁶) partially suppressed the cold-sensitive lethality of dis2-11, whereas a phosphomimetic spt5- $(T1E)_7$ mutation (which did not affect growth on its own²⁶) exacerbated it (Fig. 3e, Extended Data Fig. 8g).

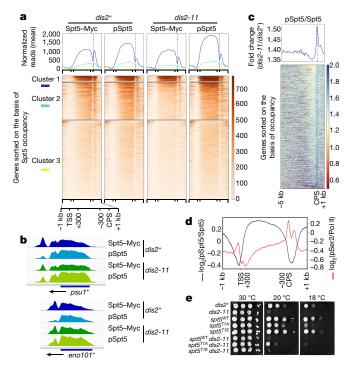


Fig. 3 | Dis2-dependent loss of pSpt5 downstream of the CPS. a, Genome-wide distribution of Spt5 and pSpt5 on transcribed genes. Metagene analyses (top) and heat maps (bottom) of ChIP-seq data reveal patterns of Spt5–Myc and pSpt5 occupancy (n = 3054 Pol II-transcribed genes) in dis2⁺ and dis2-11 cells. Each condition was analysed in duplicate. Genes were sorted on the basis of Spt5 occupancy and partitioned into three different clusters with nearest mean occupancy using k-means clustering. Note, the regions between +300 bp relative to the transcription start site (TSS) and -300 bp relative to the CPS were scaled to allow comparisons among genes of different lengths. b, Spt5 dephosphorylation downstream of the CPS. Individual gene tracks show accumulation of Spt5–Myc downstream of CPS, with lower levels of pSpt5 in *dis2*⁺ cells; this drop is attenuated in Dis2-deficient (*dis2-11*) cells. **c**, Stabilization of Spt5 phosphorylation upon inactivation of Dis2 is centred around CPS. Metagene (top) and heat map (bottom) analyses show fold-change of pSpt5:Spt5 ratio across Pol II-transcribed genes (n = 3054) in *dis2-11* versus dis2+ cells. d, Distribution of pSpt5 and Pol II-pSer2 near CPS is inversely correlated. Metagene plots of log₂(pSpt5:Spt5) and log₂(pSer2:Pol II) ratios reveal reciprocal distribution of pSpt5 and Pol II-pSer2 on transcribed genes. e, Suppression of dis2-11 by an Spt5 CTD mutant. Serial dilutions of different strains (genotypes indicated at left) grown at 30 °C, 20 °C and 18 °C. The mutant alleles tested were: spt5-(WT)₇, spt5-(T1A)₇ and spt5- $(T1E)_7$. The experiment was repeated three times independently with similar results.

We performed PRO-seq analysis⁶ to uncover the effects of Dis2 impairment on the distribution of transcribing Pol II (Fig. 4, Extended Data Fig. 9a). On individual genes, PRO-seq reads decreased within a narrowly defined zone following the CPS in *dis2*⁺ cells, but this zone extended approximately 500 bp further downstream in dis2-11 cells. Alignment of PRO-seq and ChIP-seq read distributions suggested correlation between zones of Dis2-dependent termination and Spt5 dephosphorylation (Fig. 4a). Metagene analysis of PRO-seq data revealed increased transcription beyond the CPS, both in absolute terms and relative to transcription of upstream regions, in the $\emph{dis}2$ -11 mutant (Fig. 4b, c, Extended Data Fig. 9b, c). This defect was apparent at both 18 °C and 30 °C (and in both cdk9+ and mock-treated cdk9as cells), indicating that the effects of Dis2 impairment on Pol II distribution are constitutive. Loss of viability occurs only at low temperatures, however, perhaps suggesting heightened dependence on elongation control under conditions of cold stress (although Dis2 functions in cell division might also become essential only at low temperature⁸).

To quantify termination defects, we defined two metrics on the basis of the PRO-seq read distribution: the termination index, which is

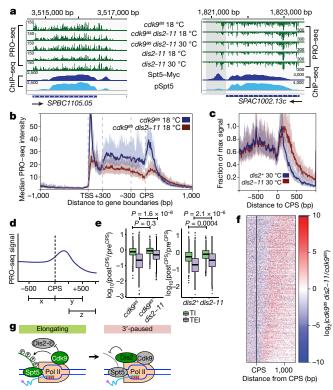


Fig. 4 | Loss of Dis2 function impairs termination. a, Transcription beyond normal termination zone in *dis2-11* cells under multiple conditions. Representative gene browser tracks show termination zones (shaded) in *dis2*⁺ or *dis2-11* cells at 18 °C or 30 °C. Alignment with ChIP-seq tracks reveals correlation between pSpt5 loss and Dis2dependent termination. All analyses were performed in the absence of 3-MB-PP1; under this condition, PRO-seq read distributions were not significantly different between cdk9⁺ and cdk9^{as} cells¹⁴. Track values reflect the maximum displayed signal (some peaks exceed these values). **b**, Genome-wide effects on Pol II distribution owing to *dis2-11* mutation. Metagene analysis of PRO-seq read distributions reveals differences in dis2-11 relative to dis2+ cells. Note, to compare genes of different lengths, regions between +300 bp relative to the TSS and -300 bp relative to the CPS were scaled. c, CPS-centred metagene analysis comparing PRO-seq read distributions in wild-type and dis2-11 cells reveals relative increase in transcription downstream of CPS in the mutant at 30 °C. Peak heights (y axis) were scaled as a fraction of maximum signal in each condition; position along the gene (x axis) was unscaled. **d**, Two metrics of termination efficiency, shown schematically: termination index (TI), the ratio of signals in the regions 500 bp downstream and upstream of CPS (y/x) and TEI, the ratio of signals in the region 250–750 bp downstream of the CPS and the region 500 bp upstream of the CPS (z/x). **b**, **c**, Solid lines represent an averaged-data plot and shaded regions represent s.d. of the median. e, The dis2-11 mutation causes a significant increase in TEI in both cdk9as (left) and cdk9+ (right) backgrounds, and in termination index in $cdk9^+$ cells (P values calculated using Welch's two sample t-test). f, Heat maps showing change in PRO-seq read distribution owing to dis2-11 mutation. Genes were ranked by decreasing TEI in cdk9as dis2-11 at 18 °C, a measure of termination-window size. All genes in **b**-**f** were required to be active and at least 1 kb from nearest genes on the same strand to eliminate effects of nearby initiation and run-through transcription (n = 939). **a-f**, n = 2 biological replicates. **g**, A transcriptionexit network comprising Cdk9, Dis2 and Spt5. At or near the CPS, Dis2 becomes active owing to a drop in Cdk9 activity and triggers Spt5 dephosphorylation, to facilitate 3'-pausing and termination.

the signal ratio in the regions 500 bp downstream or upstream of the CPS; and the termination elongation index (TEI), the ratio in regions 250–750 bp downstream and 500 bp upstream of the CPS (Fig. 4d). The dis2-11 mutation caused statistically significant increases (Welch's two-sample t-test) in the termination index in $cdk9^+$, and in the TEI in both $cdk9^+$ and $cdk9^{as}$ cells (Fig. 4e). A heat map analysis of genes

ranked on the basis of the TEI revealed termination-zone expansion consistent with imprecise termination upon loss of Dis2 function (Fig. 4f, Extended Data Fig. 9d). We also tested the effects of other dis2 alleles on Pol II distribution. Both the $dis2\Delta$ and $dis2^{T316D}$ loss-of-function mutations produced similar termination defects (Extended Data Fig. 10a–d). Unexpectedly, so did $dis2^{T316A}$, which encodes an active enzyme refractory to negative regulation by Cdk9 (Fig. 1d), suggesting that an orderly transition from elongation to termination requires both phosphorylated and unphosphorylated forms of Dis2. Cdk9 inhibition had the opposite effect, decreasing both the termination index and the TEI¹⁴, consistent with antagonism between the kinase and phosphatase.

Recent studies suggest an ordered series of events at the 3' ends of mammalian and fission-yeast protein-coding genes: 1) Pol II pausing, leading to 2) increased pS2, which promotes 3) recruitment of cleavage factors to the Pol II CTD, 4) transcript cleavage and 5) termination by the 5'-to-3' 'torpedo' exoribonuclease XRN2/Rat1^{2,3,24,27,28}. This sequence can be initiated by blocks to transcription imposed in cis², and pS2 levels are increased in 5' gene regions by mutations that decrease the intrinsic rate of Pol II²⁹, but a physiologic trigger remains unknown. A priori, both Spt5 and PP1 are probably participants in this transition—the former as a regulator of Pol II processivity and rate^{5,24,30}, the latter as a CPF component^{21,23}. Here we place Spt5 downstream of PP1 signalling in S. pombe, in which both Spt5 and a PP1 isoform are substrates of Cdk9, which is itself a positive regulator of elongation¹⁴. We define enzyme-substrate relationships among Cdk9, Dis2 and Spt5 that recapitulate a cell-cycle regulatory module⁴ and suggest a model of transcriptional exit (Fig. 4g): Dis2-dependent dephosphorylation of Spt5 and possibly other Cdk9 targets reverses elongation-rate enhancement and facilitates Pol II capture by the torpedo.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0214-z.

Received: 9 August 2017; Accepted: 17 April 2018; Published online 13 June 2018.

- Proudfoot, N. J. Transcriptional termination in mammals: stopping the RNA polymerase II juggernaut. Science 352, aad9926 (2016).
- Davidson, L., Muniz, L. & West, S. 3' end formation of pré-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells. Genes Dev. 28, 342–356 (2014).
- Sansó, M. et al. P-TEFb regulation of transcription termination factor Xrn2 revealed by a chemical genetic screen for Cdk9 substrates. Genes Dev. 30, 117–131 (2016).
- Grallert, A. et al. A PP1–PP2A phosphatase relay controls mitotic progression. Nature 517, 94–98 (2015).
- Shetty, A. et al. Spt5 plays vital roles in the control of sense and antisense transcription elongation. Mol. Cell 66, 77–88.e75 (2017).
- Booth, G. T., Wang, I. X., Cheung, V. G. & Lis, J. T. Divergence of a conserved elongation factor and transcription regulation in budding and fission yeast. *Genome Res.* 26, 799–811 (2016).
- Blethrow, J. D., Glavy, J. S., Morgan, D. O. & Shokat, K. M. Covalent capture of kinase-specific phosphopeptides reveals Cdk1-cyclin B substrates. *Proc. Natl Acad. Sci. USA* 105, 1442–1447 (2008).
- Ohkura, H., Kinoshita, N., Miyatani, S., Toda, T. & Yanagida, M. The fission yeast dis2+ gene required for chromosome disjoining encodes one of two putative type 1 protein phosphatases. Cell 57, 997–1007 (1989).
- Yamano, H., Ishii, K. & Yanagida, M. Phosphorylation of dis2 protein phosphatase at the C-terminal cdc2 consensus and its potential role in cell cycle regulation. EMBO J. 13, 5310–5318 (1994).
- Kinoshita, N., Ohkura, H. & Yanagida, M. Distinct, essential roles of type 1 and 2A protein phosphatases in the control of the fission yeast cell division cycle. *Cell* 63, 405–415 (1990).
- Viladevall, L. et al. TFIIH and P-TEFb coordinate transcription with capping enzyme recruitment at specific genes in fission yeast. *Mol. Cell* 33, 738–751 (2009).
- Pei, Y. & Shuman, S. Characterization of the Schizosaccharomyces pombe Cdk9/ Pch1 protein kinase: Spt5 phosphorylation, autophosphorylation, and mutational analysis. J. Biol. Chem. 278, 43346–43356 (2003).
- Sansó, M. et al. A positive feedback loop links opposing functions of P-TEFb/ Cdk9 and histone H2B ubiquitylation to regulate transcript elongation in fission yeast. PLoS Genet. 8, e1002822 (2012).



- Booth, G. T., Parua, P. K., Sansó, M., Fisher, R. P. & Lis, J. T. Cdk9 regulates a promoter-proximal checkpoint to modulate RNA polymerase II elongation rate in fission yeast. *Nat. Commun.* 9, 543 (2018).
- Sajiki, K. et al. Genetic control of cellular quiescence in S. pombe. J. Cell Sci. 122, 1418–1429 (2009).
- Cho, E. J., Kobor, M. S., Kim, M., Greenblatt, J. & Buratowski, S. Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. Genes Dev. 15, 3319–3329 (2001).
- Hausmann, S. & Shuman, S. Characterization of the CTD phosphatase Fcp1 from fission yeast. Preferential dephosphorylation of serine 2 versus serine 5. J. Biol. Chem. 277, 21213–21220 (2002).
- Schwer, B., Ghosh, A., Sanchez, A. M., Lima, C. D. & Shuman, S. Genetic and structural analysis of the essential fission yeast RNA polymerase II CTD phosphatase Fcp1. RNA 21, 1135–1146 (2015).
- St. Amour, C. V. et al. Separate domains of fission yeast Cdk9 (P-TEFb) are required for capping enzyme recruitment and primed (Ser7-phosphorylated) Rpb1 carboxyl-terminal domain substrate recognition. *Mol. Cell. Biol.* 32, 2372–2383 (2012).
- Pei, Y. et al. Cyclin-dependent kinase 9 (Cdk9) of fission yeast is activated by the CDK-activating kinase Csk1, overlaps functionally with the TFIIH-associated kinase Mcs6, and associates with the mRNA cap methyltransferase Pcm1 in vivo. Mol. Cell. Biol. 26, 777–788 (2006).
- Vanoosthuyse, V. et al. CPF-associated phosphatase activity opposes condensin-mediated chromosome condensation. *PLoS Genet.* 10, e1004415 (2014).
- Nedea, E. et al. The Glc7 phosphatase subunit of the cleavage and polyadenylation factor is essential for transcription termination on snoRNA genes. Mol. Cell 29, 577–587 (2008).
- Schreieck, A. et al. RNA polymerase II termination involves C-terminal-domain tyrosine dephosphorylation by CPF subunit Glc7. *Nat. Struct. Mol. Biol.* 21, 175–179 (2014).
- Baejen, C. et al. Genome-wide analysis of RNA polymerase II termination at protein-coding genes. Mol Cell 66, 38–49.e36 (2017).
- Bernecky, C., Herzog, F., Baumeister, W., Plitzko, J. M. & Cramer, P. Structure of transcribing mammalian RNA polymerase II. Nature 529, 551–554 (2016).
- Schneider, S., Pei, Y., Shuman, S. & Schwer, B. Separable functions of the fission yeast Spt5 carboxyl-terminal domain (CTD) in capping enzyme binding and transcription elongation overlap with those of the RNA polymerase II CTD. Mol. Cell. Biol. 30, 2353–2364 (2010).
- Fong, N. et al. Effects of transcription elongation rate and Xrn2 exonuclease activity on RNA polymerase II termination suggest widespread kinetic competition. *Mol. Cell* 60, 256–267 (2015).
- Glover-Cutter, K., Kim, S., Espinosa, J. & Bentley, D. L. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat. Struct. Mol. Biol.* 15, 71–78 (2008).

- Fong, N., Saldi, T., Sheridan, R. M., Cortazar, M. A. & Bentley, D. L. RNA Pol II dynamics modulate co-transcriptional chromatin modification, CTD phosphorylation, and transcriptional direction. *Mol. Cell* 66, 546–557.e3 (2017).
- Yamada, T. et al. P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. *Mol. Cell* 21, 227–237 (2006).

Acknowledgements R.P.F. is grateful for the mentorship provided by Günter Blobel (1936–2018). We thank I. M. Hagan, B. Schwer, S. Shuman, V. Vanoosthuyse, M. Yanagida, M. J. O'Connell and the National BioResource Project/Yeast Genetic Resource Center for providing yeast strains and/or antibodies; K. M. Shokat for providing 3-MB-PP1; C. Zhang for guidance in analogue-sensitive allele optimization; D. Hasson for advice on ChIP-seq data analysis and N. Steinbach and R. Parsons for assistance in phosphatase-activity measurements. J.C.T. was supported by Canadian Institutes of Health Research grant MOP-130362 and by a fellowship from Fond de recherche Quebec Santé (3315). This work was supported by National Institutes of Health grants GM25232 to G.T.B. and J.T.L. and GM104291 to R.P.F. Next-generation sequencing was supported in part by grant P30 CA196521 to the Tisch Cancer Institute

Author contributions P.K.P. and M.S. identified PP1 isoforms as Cdk9 substrates. P.K.P. conducted enzymologic studies, measured Spt5 dephosphorylation rates in wild-type and PP1 mutant backgrounds, performed ChIP-qPCR analysis and characterized genetic interactions between *dis2* and *spt5* mutant alleles. P.K.P., M.S., B.B. and J.C.T. performed ChIP-seq analysis to map distribution of pSpt5. G.T.B. performed PRO-seq analysis and developed metrics to quantify termination defects in *dis2* mutants. P.K.P., G.T.B., J.T.L. and R.P.F. prepared the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0214-z.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0214-z.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.P.F. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized The investigators were not blinded to allocation during experiments and outcome assessment.

Yeast strains and standard methods. Fission-yeast strains used in this study are listed in Supplementary Table 1^{31,32}. New strains were generated by standard techniques³³. Cells were grown in YES medium at 30 °C unless otherwise specified.

Immunological methods. Antibodies used in this study recognized pSpt5 or total Spt5¹³, Dis2-pT316⁴ or total Dis2³⁴, Myc epitope (EMD Millipore, 05-724), total Pol II (BioLegend, MMS-126R), Pol II pSer2 (Abcam, ab5095), Pol II pTyr1 (Active Motif, 61383, clone 3D12), α-tubulin (Sigma, T-5168) or GFP (Invitrogen, rabbit polyclonal (A11122) or Santa Cruz, mouse monoclonal (sc-9996)). Proteins were visualized in immunoblot analysis using either enhanced chemiluminescence (ECL, HyGLO HRP detection kit, Denville Scientific, E250) or with the Odyssey Imaging System (LI-COR Biosciences).

Kinase and phosphatase assays. Kinase assays were performed with purified proteins (Cdk9–Pch1 complex and GST–Spt5(801–990)) and $[\gamma^{-32}P]$ ATP (PerkinElmer, BLU002A250UC), as described previously²⁰. GST-PP1 was expressed in Escherichia coli at 16 °C for 16 h and purified with Glutathione Sepharose 4 Fast Flow beads³⁵. To measure protein phosphatase activity, PP1 isoforms (GST-Dis2 or GST-Sds21) purified from E. coli (2 μg) or immunoprecipitated from fission-yeast extracts (4 mg total protein) were incubated with 50 μM peptide (Spt5-NP, Spt5-pT1 or H3pS10) at 37 °C for 1 h. Colorimetric assays were performed in triplicate using BioMOL Green (Enzo Life Sciences, BML-AK111) in 25 mM HEPES, pH 7.5, 100 mM NaCl, 1 mM MnCl₂ and 1 mM DTT, in 96-well plates as described in the manufacturer's protocol. To test Dis2 activity after phosphorylation by Cdk9, Dis2 immunoprecipitated from fission-yeast extract (8 mg total protein) was incubated with Cdk9 (activated by Csk1) or mock-treated (no kinase added) in kinase assay buffer (25 mM HEPES, pH 7.5, 10 mM MgCl₂, 1 mM ATP, 1 mM DTT) for 1 h at 25 °C, washed three times with phosphatase assay buffer and tested for activity as described above.

ChIP analysis. Immunoprecipitation and ChIP were carried out using published methods^{19,36,37}. qPCR was performed with USB VeriQuest SYBR Green qPCR Master Mix (2×) (Affymetrix, 75600) in 384-well plates. ChIP was performed with three or four biological replicates and qPCR was performed in triplicate for each sample. Oligonucleotides used as primers are listed in Supplementary Table 2. In ChIP-qPCR calculations, input Ct values were first corrected by dilution factor (Raw $C_{\rm t}^{\rm input} - \log_2({\rm dilution~factor})$; dilution factor = 10 or 100), ChIP signals were normalized against input ($\Delta C_t = C_t^{\text{IP}} - C_t^{\text{input}}$), and ChIP yield was expressed as percentage input and calculated by the equation: $2^{-\Delta Ct} \times 100\%$. Fold enrichment of specific (S) ChIP-signal over nonspecific (NS) control (rabbit IgG or ChIP signals from untagged strains) was calculated by the equation: $2^{-\Delta\Delta C_t}$ in which $\Delta\Delta C_t = \Delta C_t^S - \Delta C_t^{NS}$. To measure dependency of Dis2-T316 phosphorylation on CDKs in vivo, cells expressing GFP-Dis2 from the chromosomal dis2 locus in wild-type, cdk9as, mcs6as5 or lsk1as backgrounds were grown at 30 °C to a density of approximately 1.5 \times 10⁷ cells per ml, treated with either DMSO or 10 μ M 3-MB-PP1 for 10 min, and crosslinked with 1% (v/v) formaldehyde for 15 min at 25 °C. Chromatin extracts were prepared according to a standard protocol for ChIP sample preparation ^{19,36,37} and 5 mg total protein was subjected to immunoprecipitation with anti-GFP antibody (sc-9996) and immunoblot analysis with either anti-GFP or anti-Dis2-pT316 antibody⁴.

Chemical genetics. To measure rates of Spt5 and Rpb1 dephosphorylation after CDK inhibition, cells were grown at 30 °C in YES medium to a density of approximately 1.2×10^7 cells per ml, collected by centrifugation at 25 °C, resuspended in fresh YES (preincubated at desired temperature) and incubated for 10 min before addition of DMSO or 3-MB-PP1 (10 or 20 μM). At intervals thereafter, cells (approximately 0.6×10^8) were transferred directly to tubes containing $500\,\mu l\,100\%$ (w/v) trichloroacetic acid (TCA) and collected by centrifugation. Protein extracts were prepared in the presence of 20% (w/v) TCA 38 and processed for immunoblot analysis with appropriate antibodies.

ChIP-seq. The dis2⁺ or dis2-11 cells were grown at 30 °C before crosslinking in 1% (v/v) formaldehyde. ChIP was performed with 250 µg of total protein and 1 µg of antibody. Multiplexed ChIP-seq libraries were prepared using the NEBNext Ultra II DNA Library Preparation kit (E7103S) with 25 ng of input or immunoprecipitated DNA and barcode adaptors (NEBNext Multiplex Oligos for Illumina (Set 1, E7335 and Set 2, E7500)). Paired-end sequencing (50-nt reads) was performed on an Illumina NextSeq 500. Quality control and adaptor trimming of FASTQ (raw sequencing) files were done in Galaxy using FastQC Read Quality reports (Galaxy v.0.69) and trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy v.0.36.3), respectively. Processed FASTQ files were aligned to the S. pombe genome using Bowtie2³⁹ (Galaxy v.2.2.6.2). Aligned sequences of each biological replicate were fed into MACS2⁴⁰ (Galaxy v.2.1.1.20160309.0) to call peaks from alignment results. Generated 'bedgraph treatment' files were converted to bigwig using 'Wig/BedGraph-to-bigWig converter' (Galaxy v.1.1.0), concatenated

(Galaxy v.1.0.1) to combine replicates of each sample, converted to bigwig using 'Wig/BedGraph-to-bigWig converter' (Galaxy v.1.1.0) and processed using computeMatrix (Galaxy v.2.3.6.0) in DeepTools⁴¹ to prepare data for plotting heat maps and/or profiles of given regions. Heat map and metagene plots were generated using 'plotHeatmap' (Galaxy v.2.5.0.0) and 'plotProfile' (Galaxy v.2.5.0.0) tools, respectively. The pSpt5:Spt5 signal ratios and fold change of the signal ratios in dis2-11 versus $dis2^+$ were calculated using bigwigCompare (Galaxy v.2.5.0.0). First, all ChIP-seq signals were normalized by subtracting non-specific IgG signals generated in each strain. Next, pSpt5 signals were normalized over Spt5-Myc signals in corresponding strains. Finally, the fold-change of pSpt5/Spt5 in dis2-11 versus $dis2^+$ cells was calculated using bigwigCompare (Galaxy v.2.5.0.0). For metagene analysis, all Pol II active and filtered genes were used $(n=3,122)^6$.

PRO–seq. Two batches of PRO–seq⁴² experiments were performed, each with slight differences in run-on procedure and library preparations. For each strain, biological replicates were derived from separately picked colonies. Cultures were grown in YES medium at 30 °C overnight and diluted to an $OD_{600\,\mathrm{nm}}=0.2$. After reaching approximately $OD_{600\,\mathrm{nm}}=0.5$, an equal number of cells (on the basis of OD) was set aside for all treatments (approximately 10 ml culture). At this point, a fixed amount of thawed *S. cerevisiae* (50 µl, OD = 0.68) was spiked into each sample. Cultures were then immediately spun down at 4 °C and subjected to permeabilization and library preparation. In the batch of experiments with temperature shifts, cells were spun down and resuspended in fresh YES medium, preconditioned at the desired temperature (30 °C or 18 °C) and allowed to incubate for 10 min before permeabilization and library preparation.

Samples from the temperature-varied experiments were prepared according to the standard procedure, which uses each of the four biotinylated NTPs during the run-on reaction as well as previously described adaptors⁴³. For the batch of experiments lacking temperature shifts (Extended Data Fig. 10), precision run-on reactions were performed with only two biotinylated NTPs (biotin-11-C and biotin-11-U) in addition to an equal concentration of rATP and rGTP (the two-biotin run-on approach saves costs but modestly reduces the resolution of the assay). Additionally, for these libraries a novel 3' RNA adaptor was used. Here, an adaptor with a distinct hexanucleotide sequence preceded by a 5'-monophosphorylated guanine at the 5' end and with a deoxythymidine linked 3'-3' at the 3' end is used for each library prepared (5' 5Phos-GNNN NNNGAUCGUCGGACUGUAGAACUCUGAAC-inverted dT). This results in a distinct barcode for each library and permits the pooling of all samples after the 3' end ligation step. After pooling, the remainder of the library preparation was carried out as previously described⁴³, but in a single tube. After sequencing, the inline barcode was used to parse reads on the basis of their sample of origin. Reads were processed and aligned for analysis as described14.

Experimental batches. With the exception of wild-type data used in Fig. 4 (which were taken from previously published work ¹⁴), two batches of experiments were performed to generate the precision run-on data in this work. To minimize noise introduced from across-batch comparisons, all analyses were restricted to PRO-seq libraries prepared within the same batch. However, several features of the second batch of samples (used in Extended Data Fig. 10) limited our further use of these data: 1) there were subtle profile differences that complicated direct comparisons between batches of experiments, possibly owing to the different run-on procedure described above; 2) raw read length was much shorter in the second batch, producing fewer uniquely aligning reads, possibly reflecting lower quality of the preparations; and 3) normalization that was based on spike-in did not reproducibly capture changes in transcription profiles, thus limiting direct quantitative comparisons between strains. Nonetheless, results using these samples clearly and independently recapitulated the Dis2-dependent influence on termination described in Fig. 4.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

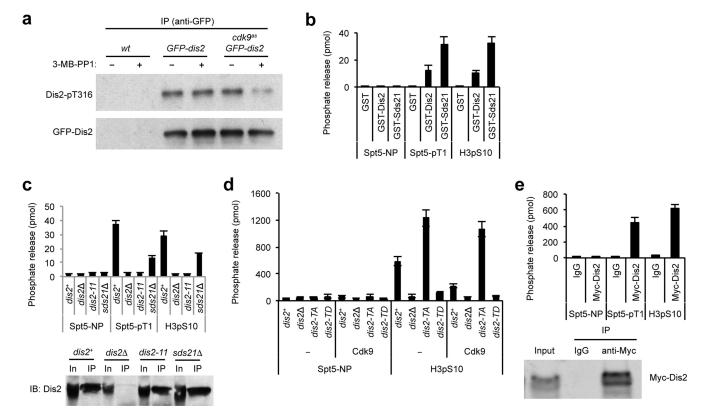
Code availability. Custom scripts and alignment pipelines have been made publicly available through a GitHub repository: https://github.com/gregtbooth/Pombe_PROseq.

Data availability. Source Data for results plotted in Figs. 1d, 2a, c-e, 4, and in Extended Data Figs. 1b-e, 3a-d, 4a-d, 5a-d, 6a, b, e, 8g are available online. The raw and processed sequencing files have been submitted to the NCBI Gene Expression Omnibus (GEO) under accession number GSE102590.

- 31. Hayashi, A. et al. Localization of gene products using a chromosomally tagged GFP-fusion library in the fission yeast *Schizosaccharomyces pombe*. *Genes Cells* **14**, 217–225 (2009).
- 32. Saiz, J. E. & Fisher, R. P. A CDK-activating kinase network is required in cell cycle control and transcription in fission yeast. *Curr. Biol.* **12**, 1100–1105 (2002).
- Moreno, S., Klar, A. & Nurse, P. Molecular genetic analysis of fission yeast Schizosaccharomyces pombe. Methods Enzymol. 194, 795–823 (1991).
- Stone, E. M., Yamano, H., Kinoshita, N. & Yanagida, M. Mitotic regulation of protein phosphatases by the fission yeast sds22 protein. *Curr. Biol.* 3, 13–26 (1993).

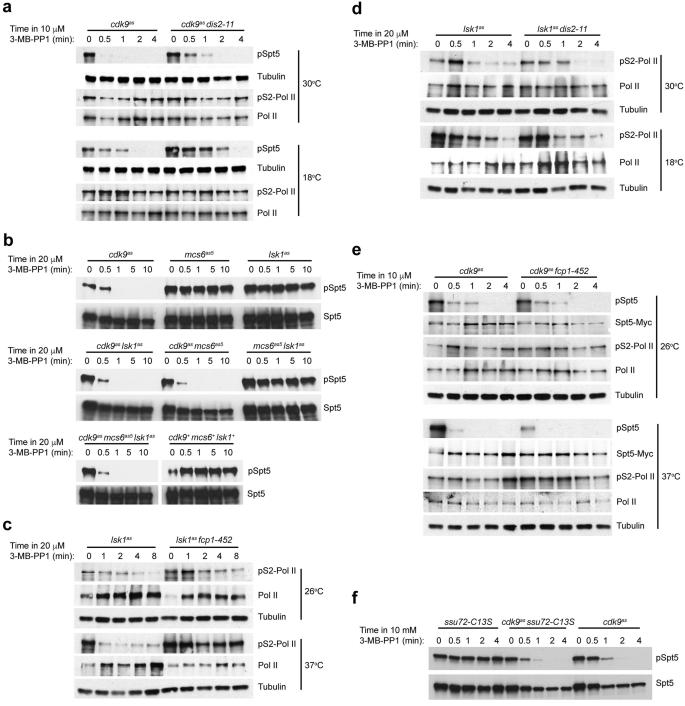


- 35. Parua, P. K., Mondal, A. & Parrack, P. HflD, an *Escherichia coli* protein involved in the λ lysis–lysogeny switch, impairs transcription activation by $\lambda \text{CII.}$ Arch. Biochem. Biophys. 493, 175-183 (2010).
- 36. Sansó, M. et al. Gcn5 facilitates Pol II progression, rather than recruitment to nucleosome-depleted stress promoters, in Schizosaccharomyces pombe. Nucleic Acids Res. 39, 6369-6379 (2011).
- 37. Tanny, J. C., Erdjument-Bromage, H., Tempst, P. & Allis, C. D. Ubiquitylation of histone H2B controls RNA polymerase II transcription elongation independently of histone H3 methylation. Genes Dev. 21, 835-847 (2007).
- 38. Kao, C. F. & Osley, M. A. In vivo assays to study histone ubiquitylation. Methods 31, 59-66 (2003).
- Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
 Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. Nat. Protocols 7, 1728–1740 (2012).
- 41. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 44, W160-W165 (2016).
- 42. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science 339, 950–953 (2013).
- 43. Mahat, D. B. et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat. Protocols 11, 1455-1476 (2016).



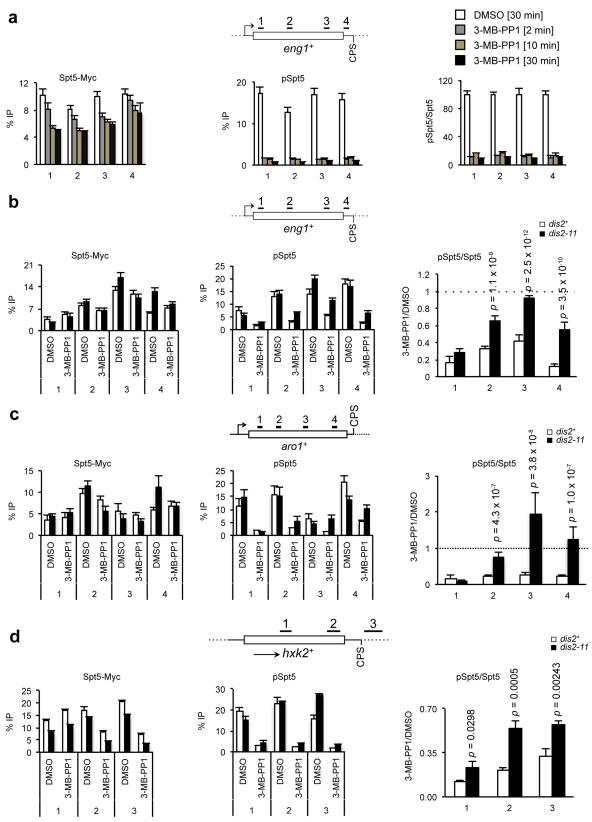
Extended Data Fig. 1 | A Cdk9-Dis2-Spt5 regulatory circuit. a, Cdk9-dependence of Dis2-T316 phosphorylation in vivo. Cells of wild-type (wt) or $cdk9^{as}$ strains, with or without GFP-tagged Dis2 expressed from the chromosomal $dis2^+$ locus, were treated for 10 min with 20 µM 3-MB-PP1 or mock-treated, as indicated. Chromatin extracts were immunoprecipitated with anti-GFP antibodies and probed with antibodies specific for Dis2 phosphorylated at Thr316 (Dis2-pT316) or GFP (n=2 independent repeats). b, Spt5 dephosphorylation by purified PP1 in vitro. Purified GST-Dis2 and GST-Sds21 were incubated with a control phosphopeptide derived from histone H3 (H3pS10), an Spt5 CTD consensus phosphopeptide (Spt5-pT1) or a non-phosphorylated peptide of the same sequence (Spt5-NP). c, Spt5 dephosphorylation by PP1 isolated

from fission yeast. A polyclonal anti-Dis2 antibody immunoprecipitates an active pSpt5 phosphatase from extracts of $dis2^+$ but not $dis2\Delta$ or dis2-11 mutant cells. Note, the antibody cross-reacts with Sds21 in immunoblots but does not efficiently immunoprecipitate Sds21. **d**, Loss of Dis2 activity upon Cdk9-dependent phosphorylation. As in Fig. 1d, results show activity of Dis2 (isolated from yeast and phosphorylated by Cdk9 or mock-treated) towards H3pS10. **e**, Top, anti-Myc immunoprecipitates from Myc-Dis2-expressing cells were tested for phosphatase activity towards Spt5-pT1 and H3pS10 peptides. Bottom, immunoblot to verify expression and immunoprecipitation of Myc-Dis2. **b**-**e**, Data are mean + s.d. from three biological replicates.



Extended Data Fig. 2 | Distinct kinase–phosphatase circuits regulate phosphorylation of Spt5 Thr1 and Rpb1 Ser2 in vivo. a, Rapid dephosphorylation of Spt5 after Cdk9 inhibition and stabilization of pSpt5 by Dis2 inactivation occur in the $spt5^+$ strain and do not depend on the C-terminal Myc-epitope tag. b, Spt5 dephosphorylation kinetics in single, double and triple cdk^{as} mutants treated with 3-MB-PP1 show that Cdk9 is the sole kinase needed to phosphorylate this site in vivo. c, Fcp1 inactivation stabilizes Rpb1 Ser2 phosphorylation after Lsk1 inhibition. Fission-yeast strains, $lsk1^{as}$ or $lsk1^{as}$ fcp1-452, were grown at 30 °C and shifted to 37 °C (or not shifted), treated for the indicated time with 20 μ M 3-MB-PP1, and analysed by immunoblotting for Pol II Ser2 phosphorylation. Note, CTD dephosphorylation leads to increased reactivity with 8WG16 antibody used to detect total Pol II. d, Dis2 activity

is dispensable for Pol II CTD Ser2 dephosphorylation. As in Fig. 2a, except that the experiment was performed in $lsk1^{as}$ cells, $20~\mu M$ 3-MB-PP1 was added and extracts were probed for Pol II Ser2 phosphorylation (or tubulin as a loading control). e, Fcp1 activity is dispensable for Spt5 Thr1 dephosphorylation. As in c, except that strains carried a $cdk9^{as}$ allele and were tested for both pSpt5 (unaffected by Fcp1 inactivation) and pSer2 (unaffected by Cdk9 inhibition). f, CPF-associated Pol II pSer5 phosphatase Ssu72 is dispensable for Spt5 Thr1 dephosphorylation. Fission-yeast strains $(cdk9^{as}, cdk9^{as} ssu72^{C138})$ and $ssu72^{C138})$ were grown at 30 °C, treated for the indicated time with 10 μ M 3-MB-PP1 and analysed by immunoblotting for pSpt5 and total Spt5. a–f, n=2 independent repeats.

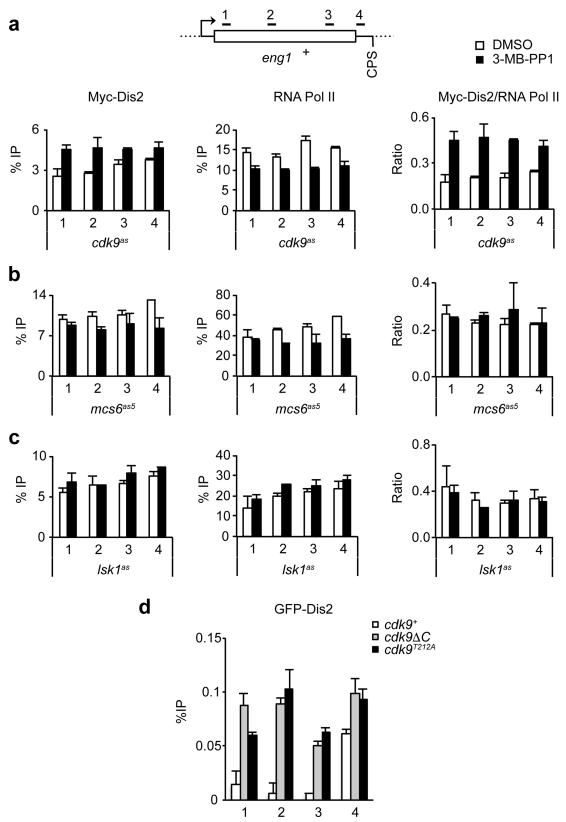


Extended Data Fig. 3 | See next page for caption.



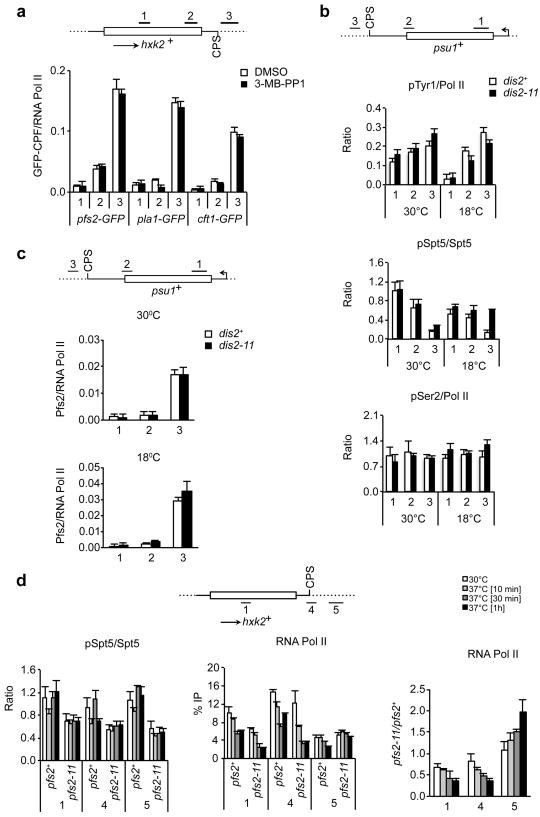
Extended Data Fig. 3 | Chromatin-associated Spt5 is dephosphorylated rapidly upon Cdk9 inhibition and stabilized in *dis2-11* cells. a, Rapid pSpt5 turnover on chromatin. ChIP-qPCR analysis of crosslinking of pSpt5 versus total Spt5 at the *eng1*⁺ gene after 3-MB-PP1 treatment for various lengths of time. Left, absolute ChIP signals for anti-Myc; middle, absolute signals for anti-pSpt5; right, ratio of pSpt5 to total Spt5, expressed as a percentage of the ratio in the absence of the inhibitor. b, Loss of Dis2 function stabilizes pSpt5 on chromatin. Left and middle, either *cdk9*^{as} *spt5-13Myc dis2*⁺ or *cdk9*^{as} *spt5-13Myc dis2-11* cells were shifted to 18 °C and treated with 10 µM 3-MB-PP1 or mock-treated with DMSO for 2 min and subjected to ChIP-qPCR analysis at the *eng1*⁺ locus for Spt5-Myc

(left) or pSpt5 (middle). Right, the pSpt5:Spt5 signal ratios of 3-MB-PP1-treated samples versus DMSO-treated samples. The pSpt5:Spt5 signal ratios between treatments (DMSO and 3-MB-PP1) were plotted for each condition (dis2+ or dis2-11). Note, higher residual levels of pSpt5 in cdk9^{as} dis2+ cells, compared to those analysed in a, may reflect less efficient dephosphorylation at 18 °C, relative to 30 °C. c, As in b, except measuring at the aro1+ gene. d, As in b, but measuring at the hxk2+ gene (raw data for pSpt5 and total Spt5 from which ratios in Fig. 2c were calculated). a-d, Data are mean + s.d. from three biological replicates. b-d, P values (Student's t-test) are indicated between wild-type (dis2+) and mutant (dis2-11) cells.



Extended Data Fig. 4 | A specific link between Cdk9 activity and Dis2 recruitment to chromatin. a, Increased recruitment of Dis2 to chromatin is a specific consequence of Cdk9 inhibition. ChIP–qPCR analysis of Myc–Dis2 (left), Pol II (middle) and Myc–Dis2:Pol II signal ratios (right) at the $eng1^+$ locus in $cdk9^{as}$ cells, expressing Myc–Dis2 from a plasmid, treated with 20 μ M 3-MB-PP1 or DMSO for 10 min at 30 °C. b, c, Mcs6 and Lsk1 activities do not influence Dis2 recruitment to chromatin. As in

a, except cells containing $mcs6^{as5}$ (**b**) or $lsk1^{as}$ (**c**) alleles were treated with 20 μ M 3-MB-PP1 for 10 min at 30 °C. In **a**–**c**, data are mean + s.d. from three biological replicates. **d**, Constitutive cdk9 loss-of-function mutations increase GFP–Dis2 recruitment to chromatin. Dis2 occupancy at the $eng1^+$ locus analysed in $cdk9^+$ cells, a $cdk9\Delta C$ mutant and a $cdk9^{T212A}$ mutant. Data are mean + s.d. from technical duplicates of one experiment.



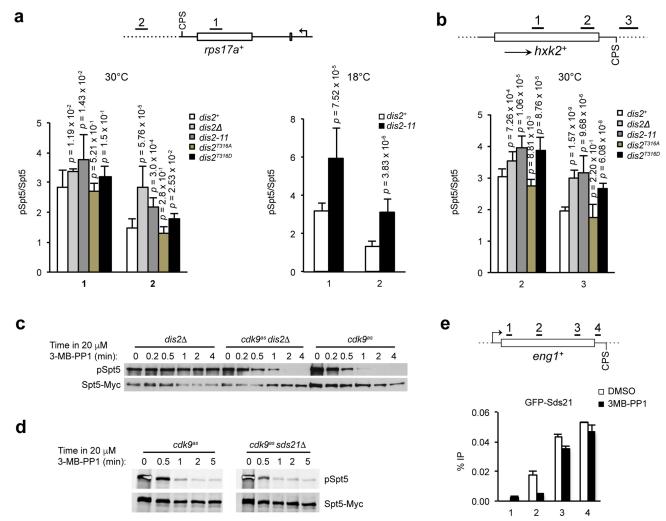
Extended Data Fig. 5 | See next page for caption.



Extended Data Fig. 5 | Regulation of pSpt5 by Cdk9 and Dis2 occurs independently of CPF recruitment and upstream of CPF function.

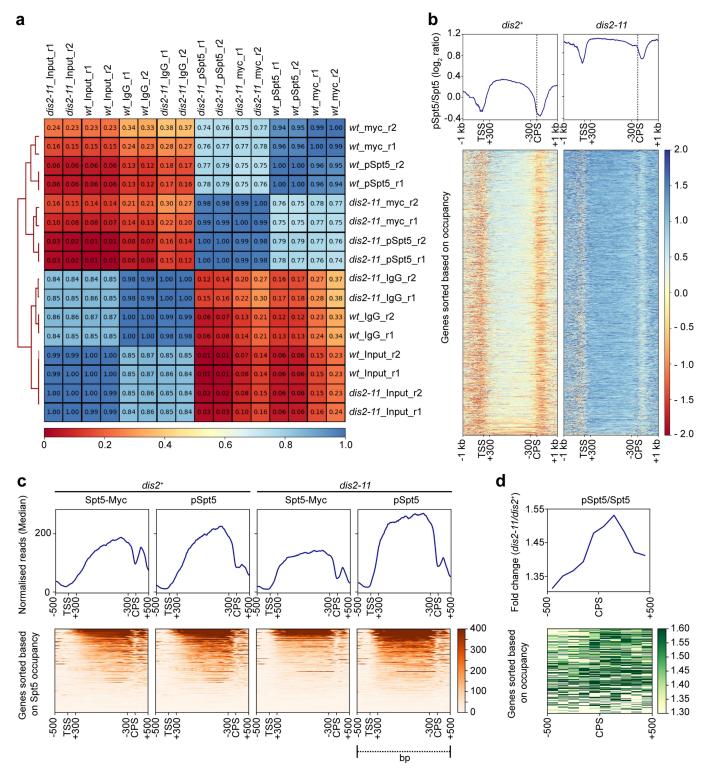
a, Core CPF recruitment to chromatin is unaffected by Cdk9 inhibition. Cells with different GFP-tagged CPF subunits expressed from their respective chromosomal loci ($cdk9^{as}$ pfs2–GFP–HA, $cdk9^{as}$ pla1–GFP–HA or $cdk9^{as}$ cft1–GFP) were grown at 30 °C and treated with 10 μ M 3-MB-PP1 or DMSO for 10 min. ChIP–qPCR analysis of GFP:Pol II signal ratios was performed at the $hxk2^+$ gene. Data are mean + s.d. from three biological replicates. b, Wild-type ($dis2^+$) and mutant (dis2-11) cells were grown at 30 °C and shifted to 18 °C (or not shifted) for 10 min. ChIP–qPCR analysis of Pol II-pTyr1:Pol II (top), pSpt5:Spt5 (middle) and Pol II-pSer2:Pol II (bottom) signal ratios was performed at the $psu1^+$ gene. c, Loss of Dis2 activity does not affect chromatin recruitment of a core

CPF subunit, Pfs2. Cells of *dis2*⁺ and *dis2-11* strains with Pfs2–GFP–HA expressed from the chromosomal *pfs2*⁺ locus were grown at 30 °C and shifted to 18 °C (or not shifted) for 10 min before formaldehyde crosslinking and chromatin isolation. ChIP–qPCR analysis of GFP:Pol II signal ratios was performed at the *psu1*⁺ gene at 30 °C (top) and 18 °C (bottom). **d**, Spt5 phosphorylation is not affected by thermal inactivation of an essential CPF subunit. Cells of *pfs2*⁺ and *pfs2-11* (temperature-sensitive) strains were grown at 30 °C, shifted to 37 °C (or not shifted) and incubated for various times as indicated. ChIP–qPCR analysis of pSpt5, total Spt5, pSpt5:Spt5 signal ratios, Pol II and Pol II signal ratios (*pfs2-11* over *pfs2*⁺), was performed at the *hxk2*⁺ locus. **b**–**d**, Data are mean + s.d. from two biological replicates.



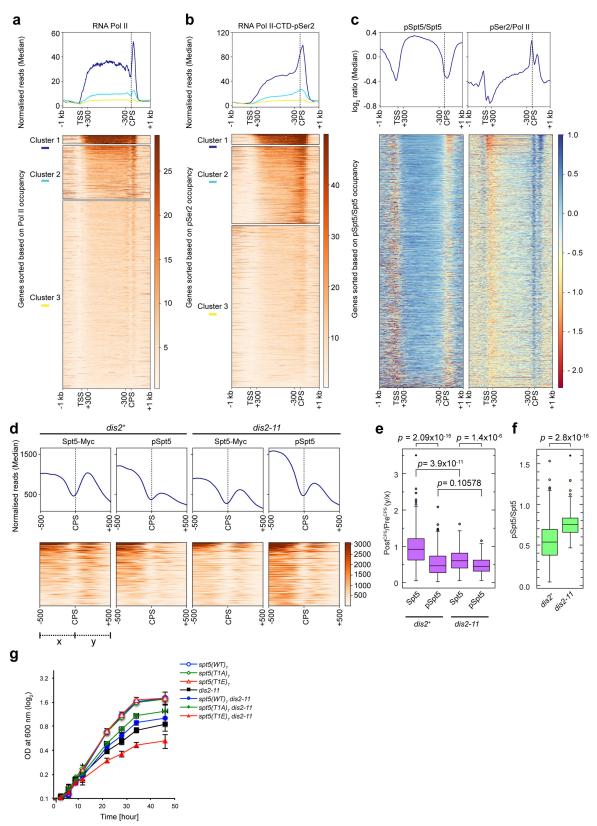
Extended Data Fig. 6 | PP1 allele- and isoform-specific stabilization of Spt5 phosphorylation on chromatin. a, ChIP–qPCR analysis at the $rps17a^+$ gene. Comparison of pSpt5:Spt5 ratio in the indicated strains upstream and downstream of the CPS at 30 °C (left) and comparison of the ratio between $dis2^+$ and $dis2^-11$ cells at 18 °C (right). b, ChIP–qPCR analysis at the $hxk2^+$ gene. Comparison of the pSpt5:Spt5 ratio in the indicated strains upstream and downstream of the CPS at 30 °C. a, b, Data are mean + s.d. from three biological replicates; P values (Student's t-test) between wild-type ($dis2^+$) and mutants ($dis2\Delta$, $dis2^-11$,

 $dis2^{T316A}$ or $dis2^{T316D}$) are indicated. c, Dephosphorylation of Spt5 after Cdk9 inhibition is retarded in a $dis2\Delta$ strain, relative to a $dis2^+$ strain. d, Spt5-dephosphorylation kinetics after Cdk9 inhibition are unaffected by sds21 deletion in a $dis2^+$ strain. c, d, n=2 independent repeats. e, Cdk9 does not restrict chromatin recruitment of Sds21. Anti-GFP ChIP–qPCR analysis at the $eng1^+$ locus in a $cdk9^{as}$ GFP–sds21 strain treated for 10 min with 10 μM 3-MB-PP1 reveals unchanged (or slightly decreased) Sds21 occupancy when Cdk9 is inhibited. Data are mean + s.d. from technical duplicates of one experiment.



Extended Data Fig. 7 | Spt5 and pSpt5 ChIP-seq analysis. a, Correlation between ChIP-seq samples. Paired-end sequencing reads were mapped to the fission-yeast genome using Bowtie2 (Galaxy v.2.2.6.2). Mapped reads of each biological replicate were used to calculate correlation between pairs of replicates. Values in boxes represent Pearson's correlation coefficients between corresponding samples (n=2 biological replicates). b, Metagene (top) and heat map (bottom) analyses show genome-wide (n=3,054 genes) comparison of pSpt5:Spt5 ratios (log₂) between $dis2^+$ and $dis2^-11$ cells (raw data from which fold change in Fig. 3c

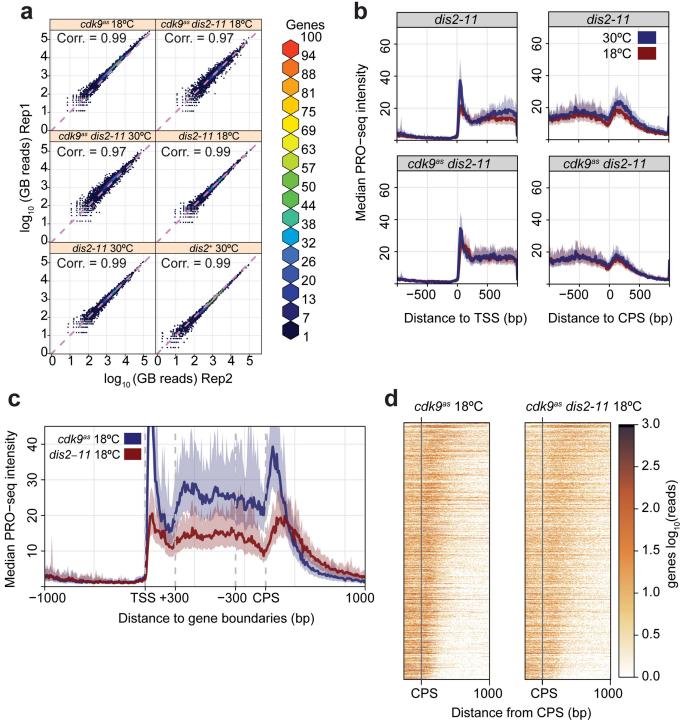
was calculated). **c**, Metagene plots (top) and heat maps (bottom) show Spt5–Myc and pSpt5 distribution in $dis2^+$ and $dis2^-11$ cells, as indicated, across Pol II-transcribed genes (n=175), filtered to include only genes separated from nearest neighbours by more than 500 bp at both ends, on both strands. **d**, Metagene plot (top) and heat map (bottom) represent fold-change of pSpt5:Spt5 ratio in $dis2^-11$ over $dis2^+$ around CPS of the genes analysed in **c**. In **b**, **c**, regions between +300 bp relative to the TSS and -300 bp relative to the CPS were scaled to enable comparison of genes of different lengths.



Extended Data Fig. 8 | See next page for caption.

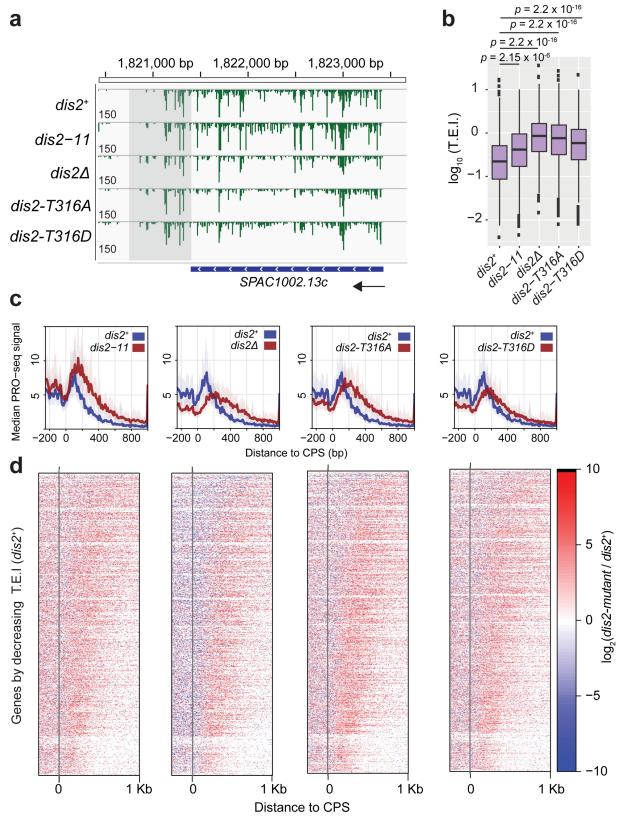
Extended Data Fig. 8 | Relative distributions of Pol II and Spt5. a, Pol II distribution on chromatin. Metagene plot (top) and heat map (bottom) of Pol II ChIP-seq distributions (data from previously published work⁵) across Pol II-transcribed genes (n = 3,054) in wild-type cells. Genes were sorted on the basis of Pol II occupancy and k-means clustering was performed to partition genes into three clusters with nearest median occupancy. **b**, Pol II pSer2 distribution on chromatin. Metagene plot (top) and heat map (bottom) of Pol II pSer2 (data from previously published work⁵) across Pol II-transcribed genes (n = 3,054) in wild-type cells. c, Distribution of pSpt5 and pSer2 on chromatin. Metagene analyses (top) and heat maps (bottom) show genome-wide comparison between \log_2 ratios of pSpt5:Spt5 and pSer2:Pol II for Pol II-transcribed genes (n = 3054) in wild-type cells (separate plots of data superimposed in Fig. 3d). In \mathbf{a} - \mathbf{c} , regions between +300 bp relative to the TSS and -300 bp relative to the CPS were scaled to enable comparison of genes of different lengths. d, Spt5 and pSpt5 distribution around CPS of Pol II-transcribed genes. Metagene plots (top) and heat maps (bottom) represent the distribution of Spt5–Myc and pSpt5 around the CPS (-500 to +500 bp)of highly active genes (n = 137; cluster 1 of Fig. 3a in $dis2^+$ and dis2-11

cells. e, Statistical analysis of Spt5-Myc and pSpt5 occupancy around CPS. Box plots represent occupancy of Spt5 and pSpt5 in the region 500 bp downstream of the CPS (y; Post^{CPS}) versus the region 500 bp upstream of the CPS (x; Pre^{CPS}), for genes in cluster 1 of Fig. 3a (n = 137), in $dis2^+$ and dis2-11 cells (Spt5, $dis2^+$: high = 3.506, low = 5.5312 × 10⁻², median = 0.9199, 95% confidence interval; pSpt5, $dis2^+$: high = 2.085, $low = 2.8570 \times 10^{-2}, median = 0.4655, 95\% \ confidence \ interval; Spt5,$ dis2-11: high = 1.617, low = 5.7597 × 10⁻², median = 0.6042, 95% confidence interval; pSpt5, dis2-11: high = 1.155, low = 6.2344 × 10⁻², median = 0.4530, 95% confidence interval). f, Box plots represent statistical significance of increases in pSpt5:Spt5 ratios in $\overset{\frown}{Post}^{CPS}$ versus Pre^{CPS} regions in dis2-11 versus $dis2^+$ cells (n = 137) ($dis2^+$: high = 1.533, low = 4.5765×10^{-2} , median = 0.5356, 95% confidence interval; dis2-11: high = 1.601, low = 0.4651, median = 0.7482, 95% confidence interval). e, f, P values were calculated using two-sided Student's t-test. g, An Spt5 that cannot be phosphorylated suppresses conditional lethality of dis2-11. Growth kinetics in liquid culture of indicated strains after a shift to 18 °C. Data are mean \pm s.d. from two biological replicates.



Extended Data Fig. 9 | The *dis2-11* mutation affects global transcription properties independent of temperature. a, PRO–seq experiments are reproducible. Scatter plots comparing PRO–seq libraries from two biological replicates for each experiment. Values represent $\log_{10}(\text{normalized reads})$ within the gene body (TSS +200 bp to the CPS) of all filtered genes (n=3383). Colours indicate the numbers of genes represented by each point. Normalization on the basis of spike-in should centre scatter about the diagonal line x=y (magenta, dotted). Correlation values represent Spearman's rank correlation. b, Comparison of composite PRO–seq profiles of *dis2-11* mutant alone (top panels) or $cdk9^{as}$ dis2-11 (bottom panels) at 18 °C and 30 °C. Profiles are centred either on the TSS (left) or the CPS (right). Shaded areas on composite profiles represent the

12.5 and 87.5% quantiles at each position. **c**, Composite PRO–seq profiles comparing $dis2^+$ strain ($cdk9^{as}$) with dis2-11 strain, both at 18 °C. Genes were scaled to a common length by fixing the middle gene body region (TSS + 300 bp to CPS – 300 bp) to 60 windows. **b**, **c**, Solid lines represent an averaged-data plot and shaded regions represent s.d. of the median. **d**, Heat maps of spike-in normalized PRO–seq signal (\log_{10}) within 10-bp windows relative to the CPS (-250 to +1,000) for $cdk9^{as}$ (left) and $cdk9^{as}$ dis2-11 (right) strains at 18 °C. Genes were ranked by decreasing TEI in $cdk9^{as}$ dis2-11 at 18 °C, a measure of the termination-window size. Each panel represents data from filtered genes that are at least 1 kb from neighbouring genes on the same strand (n=939). In $\mathbf{a}-\mathbf{d}$, data are from two biological replicates.



Extended Data Fig. 10 | See next page for caption.

RESEARCH LETTER

Extended Data Fig. 10 | Multiple *dis2* mutations cause termination defects. a, Browser image displaying normalized PRO–seq signal at the *SPAC1002.13c* gene locus. Track values reflect the maximum displayed signal (some peaks exceed these values). b, Box plots displaying the distribution of TEI values in each strain for all filtered genes separated from same-strand neighbours by at least 1 kb (n=939). Significant differences (P values from Welch's two sample t-test) in mean TEI for each strain compared with $dis2^+$ are indicated ($dis2^+$: high = 1.2304, low = -2.4005, median = -0.6532; dis2-11: high = 1.0066, low = -2.3483, median = -0.3802; $dis2^-11$: high = 1.5563, low = -1.8325, median = -0.06695; $dis2^{T316A}$: high = 1.4314, low = -2.1004, median = -0.1176; $dis2^{T316D}$: high = 1.3424, low = -2.0934,

median = -0.2310; each box shows 25th–75th percentiles). **c**, Composite PRO–seq profiles of each dis2 mutant strain (red) compared with $dis2^+$ (blue). Profiles reflect the region from -250 bp to +1,000 bp around the CPS. Shaded areas on composite profiles represent the 12.5 and 87.5% quantiles at each position. Each panel represents data from filtered genes that are at least 1 kb from neighbouring genes on the same strand (n=939). Solid lines represent an averaged-data plot of the median. **d**, Heat maps displaying $\log_2(\text{mutant/wild-type})$ PRO–seq signal within 10-bp windows from -250 bp to +1,000 bp around the CPS for all genes used in **c** sorted by decreasing TEI values in $dis2^+$ (top to bottom). In $\mathbf{a}-\mathbf{d}$, data are from two biological replicates.



Structural basis of G-quadruplex unfolding by the DEAH/RHA helicase DHX36

Michael C. Chen^{1,2}, Ramreddy Tippana³, Natalia A. Demeshkina¹, Pierre Murat², Shankar Balasubramanian^{2,4}, Sua Myong³ & Adrian R. Ferré-D'Amaré¹*

Guanine-rich nucleic acid sequences challenge the replication, transcription, and translation machinery by spontaneously folding into G-quadruplexes, the unfolding of which requires forces greater than most polymerases can exert^{1,2}. Eukaryotic cells contain numerous helicases that can unfold G-quadruplexes³. The molecular basis of the recognition and unfolding of G-quadruplexes by helicases remains poorly understood. DHX36 (also known as RHAU and G4R1), a member of the DEAH/RHA family of helicases, binds both DNA and RNA G-quadruplexes with extremely high affinity⁴⁻⁶, is consistently found bound to G-quadruplexes in cells^{7,8} and is a major source of G-quadruplex unfolding activity in HeLa cell lysates⁶. DHX36 is a multi-functional helicase that has been implicated in G-quadruplex-mediated transcriptional and posttranscriptional regulation, and is essential for heart development, haematopoiesis, and embryogenesis in mice⁹⁻¹². Here we report the co-crystal structure of bovine DHX36 bound to a DNA with a G-quadruplex and a 3' single-stranded DNA segment. We show

that the N-terminal DHX36-specific motif folds into a DNA-binding-induced α -helix that, together with the OB-fold-like subdomain, selectively binds parallel G-quadruplexes. Comparison with unliganded and ATP-analogue-bound DHX36 structures, together with single-molecule fluorescence resonance energy transfer (FRET) analysis, suggests that G-quadruplex binding alone induces rearrangements of the helicase core; by pulling on the single-stranded DNA tail, these rearrangements drive G-quadruplex unfolding one residue at a time.

DEAH/RHA helicases share a structural core^{13–18} consisting of two RecA-like domains (RecA1 and RecA2) followed by a C-terminal domain (itself comprised of degenerate-winged-helix (WH), ratchetlike (RL), and oligonucleotide and oligosaccharide-binding-fold-like (OB) subdomains). At its N terminus, DHX36 augments the DEAH/RHA core with a glycine-rich element followed by the DHX36-specific motif (DSM; Fig. 1a, Extended Data Fig. 1). The DSM is essential for binding of DHX36 to G-quadruplexes¹⁹. We

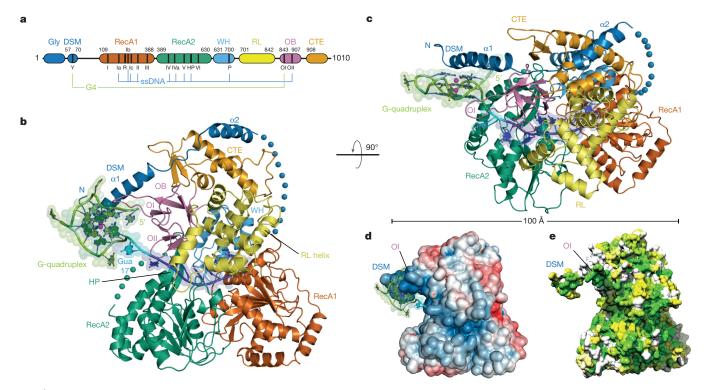


Fig. 1 | Overall structure of the DHX36–G-quadruplex DNA complex. a, Domain organization; G-quadruplex (G4)- and ssDNA-interacting regions indicated. b, Cartoon representation of the co-crystal structure of DHX36 bound to DNA Myc , colour-coded as in a. Spheres denote two disordered segments (blue, 20 and 53 residues in the crystallization

construct and wild-type, respectively; green, 13 residues). OB loops I and II (OI and OII) contact DNA. **c**, As in **b**, rotated by 90°. **d**, Electrostatic potential calculated with DNA omitted from the co-crystal structure (blue to red, $\pm 5~k_{\rm B}T$). **e**, Phylogenetic conservation among 250 DHX36 orthologues (white to green, least to most conserved).

¹Biochemistry and Biophysics Center, National Heart, Lung and Blood Institute, Bethesda, MD, USA. ²Department of Chemistry, University of Cambridge, Cambridge, UK. ³Biophysics Department, Johns Hopkins University, Baltimore, MD, USA. ⁴Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. *e-mail: adrian.ferre@nih.gov

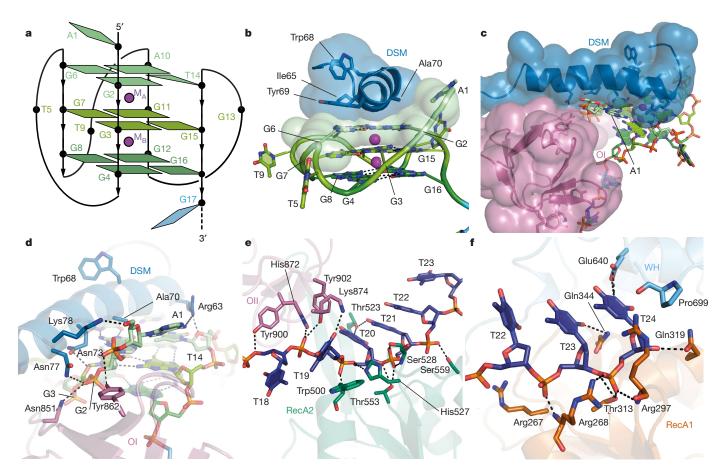


Fig. 2 | **DHX36-DNA interaction. a**, Schematic of the DHX36-bound all-parallel G-quadruplex. **b**, The DSM stacks on the 5' (top) non-canonical quartet. Transparent spheres represent van der Waals radii. **c**, The DSM and the OI loop of the OB domain flank A1. **d**, Interaction of the DSM and loop OI of DHX36 with the DNA backbone near the 5' end of DNA^{Myc}.

e, Interaction of the OII loop and the RecA2 domain with T18–T22 of the 3' single-stranded region of DNA Myc . f, Interaction of the RecA1 and WH domains of DHX36 with T23–T24 of the 3' single-stranded region of DNA Myc .

co-crystallized a DHX36 construct missing the glycine-rich element but containing the full DSM (hereafter DHX36-DSM; this construct has G-quadruplex binding and repetitive unfolding activity comparable to those of wild-type bovine and human DHX36, Extended Data Fig. 2) with a 24-nucleotide (nt) DNA (hereafter, DNA Myc) comprised of a Myc-promoter-derived G-quadruplex-forming sequence followed by a 3′ single-stranded extension of seven thymidines. We also crystallized a truncated DHX36 without the glycine-rich and DSM elements (hereafter, DHX36-core). Structures were solved through the single-wavelength anomalous dispersion (SAD) and molecular replacement methods (Extended Data Tables 1, 2, Extended Data Fig. 3; see Methods).

In the DHX36-DSM-DNA Myc complex, the RecA1, RecA2, and C-terminal domains are arranged as a trefoil (Fig. 1b). Connected to RecA1 by a disordered linker, the N-terminal extension folds into two α -helices, the first of which contains the DSM. This DSM helix projects away from the body of the helicase and contacts the 5' (top) face of the bound G-quadruplex (Fig. 1c). The OB domain contacts both the G-quadruplex and the adjacent single-stranded segment of DNA Myc, the 3'-side of which is held in a nucleic-acid-binding channel formed by the RecA1, RecA2, and C-terminal domains. The amphipathic DSM helix is overall cationic, and the path of the single-stranded DNA follows a positively charged groove between the RecA2 and C-terminal domains (Fig. 1d). This groove is too narrow to accommodate doublestranded DNA, consistent with the requirement^{6,21–23} for a 3' singlestranded extension for DHX36 activity. Phylogenetic conservation largely follows this groove and extends to the non-polar face of the DSM helix (Fig. 1e).

Solution NMR has shown^{20,24} that residues 1–17 of DNA^{Myc} fold into a stable, parallel, three-tiered G-quadruplex (Extended Data Fig. 4a, b). Our co-crystal structure reveals that association with the helicase reorganizes the DNA. Instead of three canonical G-quartets, the DHX36-bound DNA contains two G-quartets stacked underneath a non-canonical A•T•G•G quartet. The top G-quartet is absent because G17, which was the 3′-most guanine of the bottom G-quartet²⁰, has been pulled by the helicase into the 3′ single-stranded region. Shifting the DNA sequence by one residue while maintaining an overall three-tiered G-quadruplex structure with minimal propeller loops forces a new A10•T14 Watson–Crick pair to form the top quartet together with G2 and G6 (Fig. 2a, Extended Data Fig. 4b). The DHX36-bound rearranged G-quadruplex is considerably less stable than the free, canonical Myc G-quadruplex (Extended Data Fig. 5), indicative of the degree of destabilization caused by DHX36 binding alone.

The DSM has been shown ¹⁹ to be necessary but not sufficient for high-affinity G-quadruplex binding by DHX36; the full-length helicase and an isolated DSM peptide bind to G-quadruplexes with dissociation constants of below 10 pM and 310 nM, respectively ^{4,12,19}. This is consistent with the helicase core contributing to the DHX36–DNA interface (Fig. 1b, c). In the co-crystal structure, a hydrophobic core is formed by the α -helical DSM residues Ile65, Trp68, Tyr69 and Ala70, producing a flat non-polar surface that stacks on the nucleobases of the top quartet of the bound G-quadruplex (Fig. 2b), reminiscent of the mode of G-quadruplex recognition by planar small molecules ²⁵. The single-stranded A1, which is 5' to the G-quadruplex in our structure, packs between the $\alpha 1$ DSM helix and the OB domain (Fig. 2c). The C-terminal side of the $\alpha 1$ DSM helix and the first loop of the

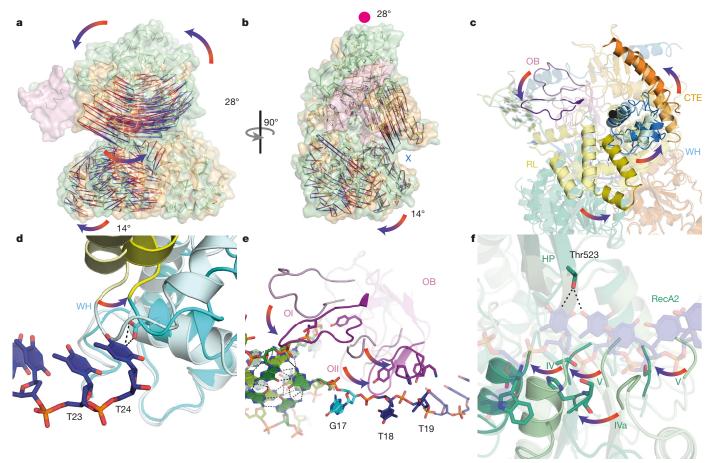


Fig. 3 | DNA-binding-induced structural transitions of DHX36. a, Superposition of DHX36-DSM-DNA Myc and DHX36-core structures (green and orange, respectively; $C\alpha$ vectors from red to blue) through RecA1. DNA Myc , pink. b, As in a, rotated by 90°. Red circle and blue cross denote C-terminal domain rotation out of and into the plane, respectively. c, C-terminal sub-domains. Unliganded and DNA-bound, pastel and solid

colours, respectively. Black circle, approximate axis of rotation. \mathbf{d} , The WH in unliganded and DNA-bound states. T24 of DNA Myc impinges on the loop linking WH and RL. \mathbf{e} , The OI and OII loops of the OB domain in unliganded and DNA-bound states. \mathbf{f} , The RecA2 domain in unliganded and DNA-bound states. Movement of conserved helicase motifs IV, IVa and V (Fig. 1a) is highlighted.

OB domain (OI loop) form extensive hydrogen bonds with the sugar-phosphate backbone of the 5′ leader (A1) and the G-quadruplex residues immediately following it (Fig. 2d). Formation of a composite G-quadruplex-binding surface between the DSM and OB domains explains why, in a previously reported NMR structure of a low-affinity complex between an 18-amino-acid DSM-derived peptide and a G-quadruplex²⁶, the DNA-binding-induced α -helix was out of register with that seen in our co-crystal structure (Extended Data Fig. 4c–g). In addition, DHX36 does not markedly discriminate between DNA and RNA substrates^{4,27}, and it recognizes the 3′ single-stranded region of DNA^{Myc} primarily by contacts with phosphates of its backbone (Fig. 2e, f). A second loop from OB (OII loop) contacts the backbones of T18 and T19 (Fig. 2e), while WH and the RecA1 domain interact with T23 and T24 (Fig. 2f).

Whereas other helicases can resolve both antiparallel and parallel G-quadruplexes²³, DHX36 has a strong preference for the latter, being inactive on fully antiparallel G-quadruplexes, and exhibiting reduced activity on G-quadruplexes with mixed parallel and antiparallel connectivity^{21,22,26}. Bound to DHX36, DNA ^{Myc} has three single-nucleotide double-chain-reversal loops (T5, T9, and G13) that do not sterically interfere with recognition of the top quartet by the DSM. Our structure suggests that the preference of DHX36 for parallel G-quadruplexes is likely to arise from the steric interference of diagonal and lateral loops with DSM binding. In addition, a 5' G-tract with the opposite polarity would interfere with binding to the OI loop.

Our DHX36-DSM-DNA Myc co-crystal structure provides an unprecedented view of the open, ATP-independent conformation adopted by

a nucleic-acid bound DEAH/RHA helicase. Superposition of the RecA1 domains of our DHX36-core and DHX36-DSM-DNA Myc structures shows that DNA binding alone induces rotations of the C-terminal and RecA2 domains by 28° and 14°, respectively (Fig. 3a-c). This conformation accommodates five stacked single-stranded (ss) DNA residues between the 5' β-hairpin (HP) and the constriction formed by Arg297, Gln319 and Pro699 (Fig. 2e, f). Compared to the ATP analogue-bound and unliganded states, the nucleic-acid-interacting elements of RecA2 (motifs IV, IVa, and V; Fig. 1a) shift away from RecA1 by 6 Å approximately the distance between successive nucleotides (Fig. 3f). HP acts as a fulcrum upon core opening, unstacking T18 and T19 on one side, and stabilizing the 3' stack of nucleotides by hydrogen bonding with Thr523 on the other (Fig. 3e, f). The opening motion may allow the G-quadruplex to unfold by one residue, and is consistent with the one-nucleotide displacement of the DHX36-bound DNA Myc structure relative to its free solution conformation (Extended Data Fig. 4a, b). Together with published structures of DEAH/RHA helicases in the ground^{16,17}, transition¹⁵, and post-hydrolysis^{13,14} states, our DHX36 structures support the hypothesis¹⁷ that DEAH/RHA helicases cycle between four- and five-nucleotide stack states enforced by the HP and a 3'-constriction site to unwind their substrates (Extended Data Figs. 6, 7).

We examined structure-guided mutants of DHX36 using a single-molecule fluorescence resonance energy transfer (smFRET) assay previously developed to characterize the repetitive, ATP-independent G-quadruplex unfolding activity of the wild-type helicase²² (Fig. 4a–c; Extended Data Fig. 2). The *Myc* promoter-derived parallel

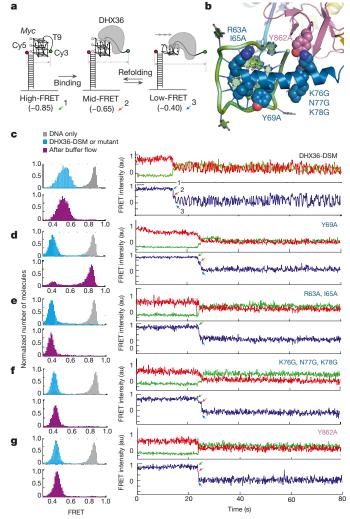


Fig. 4 | **Single-molecule FRET analysis of DHX36-DSM mutants. a**, Reporter with a G-quadruplex of the DNA ^{Myc} sequence, and a nine-thimidine single-stranded 3' tail. DHX36 shifts FRET from high (~0.8) to medium—low oscillation (~0.6, canonical DNA ^{Myc}; ~0.4, reorganized DNA ^{Myc}; Extended Data Figs. 2, 8). **b**, Structure-guided mutations. **c**, The DHX36-DSM crystallization construct remains DNA-bound upon flow and exhibits repetitive unfolding, similar to the wild type (Extended Data Fig. 2). **d**, The Y69A mutant lacks repetitive unfolding and dissociates from the G-quadruplex upon flow. **e–g**, Three mutants remain bound following flow, but lack repetitive unfolding. Each experiment was highly reproducible and in triplicate (data from more than 10,000 molecules per experiment).

G-quadruplex DNA that we use exhibits high FRET (Fig. 4a). Binding of DHX36 induces conformational changes, in which oscillations in FRET efficiency between medium and low (approximately 0.6 and 0.4, respectively) reflect repetitive unfolding between the canonical (with three complete G-quartets) and reorganized (pulled by one nucleotide) DNA Myc G-quadruplex, respectively 28 (Extended Data Fig. 8). The repetitive unfolding activity is ATP-independent, as the absence of ATP or presence of non-hydrolysable ATP analogues do not affect it (Extended Data Fig. 2i, j). We hypothesize that the repetitive unfolding activity stems from ATP-independent helicase core opening and reciprocating rotation of the C-terminal domain. ATP is likely to be required only for release of DNA from the helicase, as rapid dissociation occurs upon ATP addition (Extended Data Fig. 2i, j).

Solution NMR of a DSM-derived peptide²⁶ (Extended Data Fig. 4c–g), as well as proteolytic susceptibility of the DHX36-DSM N terminus, indicate that $\alpha 1$ is intrinsically disordered²⁹, becoming fully α -helical upon interaction with the substrate G-quadruplex. Our mutagenesis

shows that, as in other examples 29 of ligand-induced protein structure, binding free energy is distributed non-uniformly across the DSM. The R63A/I65A and KNK76GGG mutations of residues anchoring αl on the G-quadruplex backbone are less deleterious than mutation of Tyr69 (Fig. 4b, d–g). Mutation of Tyr69, which stacks directly on the top quartet (Fig. 2a), weakens the helicase–DNA association to such an extent that, uniquely among the mutants examined, this protein dissociates from DNA upon buffer flow (Fig. 4d). Ligand-induced folding of αl may allow DHX36 to mould to G-quadruplexes with different local structures, and even to antiparallel substrates with lower efficiency, during its mechanochemical cycle.

Our DHX36 co-crystal structure shows how a protein that evolved to recognize G-quadruplex-containing nucleic acids combines binding to the face and backbone of the G-quadruplex with recognition of both 5′ and 3′ single-stranded extensions. The unfolding activity of DHX36 was previously shown to be highly sensitive to the stability of its G-quadruplex substrates^{5,22}. This sensitivity is consistent with our demonstration that nucleic acid binding energy is transduced by DHX36 into a discrete, directed pulling force arising from C-terminal domain rotation and helicase core opening. These ATP-independent structural changes remodel the G-quadruplex, resulting in a substrate unwound by a single nucleotide. Our analysis thus highlights the importance of ATP-independent structural changes for nucleic acid remodelling by a canonical DEAH/RHA helicase and constitutes a starting point for further structural analysis of the mechanochemical cycle of these important enzymes.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0209-9.

Received: 23 September 2017; Accepted: 23 April 2018; Published online 13 June 2018.

- Hänsel-Hertsch, R., Di Antonio, M. & Balasubramanian, S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. Nat. Rev. Mol. Cell Biol. 18, 279–284 (2017).
- Rhodes, D. & Lipps, H. J. G-quadruplexes and their regulatory roles in biology. Nucleic Acids Res. 43, 8627–8637 (2015).
- Mendoza, O., Bourdoncle, A., Boulé, J. B., Brosh, R. M. J. Jr & Mergny, J. L. G-quadruplexes and helicases. *Nucleic Acids Res.* 44, 1989–2006 (2016)
- Giri, B. et al. G4 resolvase 1 tightly binds and unwinds unimolecular G4-DNA. Nucleic Acids Res. 39, 7161–7178 (2011).
- Chen, M. C., Murat, P., Abecassis, K., Ferré-D'Amaré, A. R. & Balasubramanian, S. Insights into the mechanism of a G-quadruplex-unwinding DEAH-box helicase. Nucleic Acids Res. 43, 2223–2231 (2015).
- Vaughn, J. P. et al. The DEXH protein product of the DHX36 gene is the major source of tetramolecular quadruplex G4-DNA resolving activity in HeLa cell lysates. J. Biol. Chem. 280, 38117–38120 (2005).
- Lattmann, S., Stadler, M. B., Vaughn, J. P., Ákman, S. A. & Nagamine, Y. The DEAH-box RNA helicase RHAU binds an intramolecular RNA G-quadruplex in TERC and associates with telomerase holoenzyme. *Nucleic Acids Res.* 39, 9390–9404 (2011).
- McRae, E. K. S. et al. Human DDX21 binds and unwinds RNA guanine quadruplexes. Nucleic Acids Res. 45, 6656–6668 (2017).
- Nie, J. et al. Post-transcriptional regulation of Nkx2–5 by RHAU in heart development. Cell Reports 13, 723–732 (2015).
- Lai, J. C. et al. The DEAH-box helicase RHAU is an essential gene and critical for mouse hematopoiesis. *Blood* 119, 4291–4300 (2012).
- Sexton, A. N. & Collins, K. The 5' guanosine tracts of human telomerase RNA are recognized by the G-quadruplex binding domain of the RNA helicase DHX36 and function to increase RNA accumulation. Mol. Cell. Biol. 31, 736–743 (2011).
- Booy, E. P. et al. The RNA helicase RHAU (DHX36) unwinds a G4-quadruplex in human telomerase RNA and promotes the formation of the P1 helix template boundary. *Nucleic Acids Res.* 40, 4110–4124 (2012).
- Walbott, H. et al. Prp43p contains a processive helicase structural architecture with a specific regulatory domain. EMBO J. 29, 2194–2204 (2010).
- He, Y., Andersen, G. R. & Nielsen, K. H. Structural basis for the function of DEAH helicases. *EMBO Rep.* 11, 180–186 (2010).
- Prabu, J. R. et al. Structure of the RNA helicase MLE reveals the molecular mechanisms for uridine specificity and RNA-ATP coupling. *Mol. Cell* 60, 487–499 (2015).
- 16. Tauchert, M. J., Fourmann, J. B., Lührmann, R. & Ficner, R. Structural insights into the mechanism of the DEAH-box RNA helicase Prp43. eLife 6, 762 (2017).



- He, Y., Staley, J. P., Andersen, G. R. & Nielsen, K. H. Structure of the DEAH/RHA ATPase Prp43p bound to RNA implicates a pair of hairpins and motif Va in translocation along RNA. RNA 23, 1110–1124 (2017).
- Chen, M. C. & Ferré-D'Amaré, A. R. Structural basis of DEAH/RHA helicase activity. Crystals 7, 253 (2017).
- Lattmann, S., Giri, B., Vaughn, J. P., Akman, S. A. & Nagamine, Y. Role of the amino terminal RHAU-specific motif in the recognition and resolution of guanine quadruplex-RNA by the DEAH-box RNA helicase RHAU. *Nucleic Acids Res.* 38, 6219–6233 (2010).
- Ambrus, A., Chen, D., Dai, J., Jones, R. A. & Yang, D. Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry* 44, 2048–2058 (2005).
- Smaldino, P. J. et al. Mutational dissection of telomeric DNA binding requirements of G4 resolvase 1 shows that G4-structure and certain 3'-tail sequences are sufficient for tight and complete binding. PLoS One 10, e0132668 (2015).
- Tippana, R., Hwang, H., Opresko, P. L., Bohr, V. A. & Myong, S. Single-molecule imaging reveals a common mechanism shared by G-quadruplex-resolving helicases. *Proc. Natl Acad. Sci. USA* 113, 8448–8453 (2016).
- Yangyuoru, P. M., Bradburn, D. A., Liu, Z., Xiao, T. S. & Russell, R. The G-quadruplex (G4) resolvase DHX36 efficiently and specifically disrupts DNA G4s via a translocation-based helicase mechanism. J. Biol. Chem. 293, 1924–1932 (2018).
- Phan, A. T., Modi, Y. S. & Patel, D. J. Propeller-type parallel-stranded G-quadruplexes in the human c-myc promoter. J. Am. Chem. Soc. 126, 8710–8716 (2004).
- Ohnmacht, S. A. & Neidle, S. Small-molecule quadruplex-targeted drug discovery. *Bioorg. Med. Chem. Lett.* 24, 2602–2612 (2014).
- Heddi, B., Cheong, V. V., Martadinata, H. & Phan, A. T. Insights into G-quadruplex specific recognition by the DEAH-box helicase RHAU: Solution structure of a peptide-quadruplex complex. *Proc. Natl Acad. Sci. USA* 112, 9608–9613 (2015).
- Creacy, S. D. et al. G4 resolvase 1 binds both DNA and RNA tetramolecular quadruplex with high affinity and is the major source of tetramolecular quadruplex G4-DNA and G4-RNA resolving activity in HeLa cell lysates. *J. Biol. Chem.* 283, 34626–34634 (2008).
- Tippana, R., Xiao, W. & Myong, S. G-quadruplex conformation and dynamics are determined by loop length and sequence. *Nucleic Acids Res.* 42, 8106–8114 (2014).

 Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 16, 18–29 (2015).

Acknowledgements We thank the staff of sector 5 of ALS and beamline 17-ID-B of APS for crystallographic data collection; Y. He, National Heart, Lung and Blood Institute (NHLBI) for protein production; G. Piszczek (NHLBI) for DSC; R. Levine and D.-Y. Lee (NHLBI) for mass spectrometry; and C. Fagan, C. Jones, T. Numata, R. Trachman III, K. Warner, and J. Zhang for discussions. This work was partly conducted at the ALS on the Berkeley Center for Structural Biology beamlines, which are supported by the US National Institutes of Health (NIH). Use of ALS and APS was supported by the US Department of Energy. This work was supported in part by the NIH (GM105453, S.M.), American Chemical Society (RSG-12-066-01-DMC, S.M.), National Science Foundation Physics Frontiers Center Program (0822613, S.M.), Wellcome Trust (099232/z/12/z, S.B.), European Research Council (339778, S.B.), Cancer Research UK (C12303/A17197 and C9681/A18618, S.B.), NIH-Oxford-Cambridge Scholars Program (M.C.C.), Cambridge Trust (M.C.C.), and the intramural program of the NHLBI, NIH.

Reviewer information *Nature* thanks D. Patel, K. Raney and V. Zakian for their contribution to the peer review of this work.

Author contributions M.C.C., P.M., S.B., and A.R.F.-D. conceived the project; M.C.C. performed protein expression, crystallization and structure determination; N.A.D. prepared mutants and characterized model G-quadruplexes; R.T. and S.M. performed smFRET; and M.C.C. and A.R.F.-D. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0209-9

Supplementary information is available for this paper at https://doi. org/10.1038/s41586-018-0209-9.

Reprints and permissions information is available at http://www.nature.com/reprints

Correspondence and requests for materials should be addressed to A.R.F.-D. **Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Protein expression and purification. Bos taurus DHX36-core, DHX36-core-SeMet, and DHX36-DSM were expressed in Escherichia coli LOBSTR (DE3)³⁰. All proteins have C-terminal 8His tags and DHX36-DSM also has an N-terminal GST tag. Starter cultures were grown at 37 °C in MDAG-135 medium³¹. Production cultures in Terrific broth were induced with 1 mM IPTG at 20 °C and grown overnight. Lysis was in 50 mM HEPES-KOH (pH 7.5), 0.5 M KCl, 10 mM β -mercaptoethanol, 0.1% (v/v) Tween-20, 10% (v/v) glycerol, and Sigmafast EDTA-free protease inhibitor cocktail. Lysate supernatant was treated with polyethyleneimine (0.05%, v/v) and loaded on a Ni-NTA Superflow (Qiagen) column. The proteins were eluted with 500 mM imidazole. DHX36-core was further purified on a Superdex 200 PG column (GE Healthcare) in 20 mM HEPES-KOH pH 7.5, 150 mM KCl, 10% (v/v) glycerol, 0.5 mM TCEP, and 2.5 mM MgCl₂. For DHX36-DSM, the eluate from the Ni-NTA column was loaded onto GSTrap 4B (GE Healthcare) column, washed with 20 mM HEPES-KOH (pH 7.5), 150 mM KCl, 10% (v/v) glycerol, and 1 mM TCEP (pH 7.0), and eluted with 25 mM reduced glutathione. The eluted DHX36-DSM was incubated at 20 °C with TEV protease (1:10 mass ratio) for 1 h. Then, DNA Myc (5'-AGG GTG GGT AGG GTG GGT TTT TTT-3') was added (2:1 DHX36-DSM:DNA Myc molar ratio) and the mixture was incubated for 30 min at 21 °C. The mixture was dialysed (50 kDa MWCO membrane) against 50 mM HEPES-KOH (pH 7.5), 150 mM KCl, and 10% (v/v) glycerol overnight at 4°C and then incubated with Amintra GST resin (Expedeon) for 1 h at 4°C. For crystallization, the complex was reductively methylated as described³². Electrospray ionization mass spectrometry (ESI-MS) indicated a mass of 109,098 \pm 2 Da, which corresponds to the dimethylation of 63 out of a total of 66 lysines in DHX36-DSM. After methylation, the DHX36-DSM-DNA Myc complex was further purified by size-exclusion chromatography as DHX36-core. For expression of DHX36-core-SeMet, PASM-5052 autoinduction expression medium³¹ was inoculated with a starter culture grown in MDAG-135 medium. Cultures were grown at 20 °C for ~6 days. Purification was as described above. The mass of DHX36-core-Semet by ESI-MS was 101,011 \pm 2 Da, corresponding to a methionine labelling efficiency of 95.8%. All proteins were purified to >98% homogeneity, with the exception of DHX36-core-SeMet, which was purified to >80% homogeneity (as judged by Coomassie blue staining of serial dilutions analysed by SDS-PAGE). All DHX36 constructs used in this study contain a deletion of residues 1-54, which encompasses the Gly-rich region. DHX36-core contains a deletion of residues 1-149. DHX36-AAA contains a KKK192AAA mutation to prevent spontaneous proteolysis. DHX36-DSM contains a deletion of residues 111–159 and surface entropy reduction mutations EEK435YYY and KDTK752AATA. All mutations used to generate the various constructs (DHX36-core, DHX36-AAA, DHX36-DSM, and structure-guided mutants) were generated using the QuikChange Lighting kit (Agilent). DHX36-DSM mutants for smFRET were purified essentially as above, but after elution from the GSTrap 4B column, the GST tag was cleaved by TEV protease in buffer with 400 mM KCl and removed by a second passage through the GSTrap 4B resin. The mutant proteins were then dialysed against 50 mM HEPES-KOH (pH 7.5), 600 mM KCl, 10% (v/v) glycerol and 1 mM TCEP (pH 7.0) overnight at 4°C and purified by size-exclusion chromatography (Superdex 200 PG, GE Healthcare) in the same buffer.

Crystallization and diffraction data collection. Hanging drops were prepared by mixing 1 µl each of DHX36-core (5 g/l) and reservoir (0.2 M ammonium citrate (pH 7.0) and 20% (w/v) PEG3350), and were equilibrated by vapour diffusion at 21 °C. DHX36-core-AlF₄⁻ was crystallized under the same conditions in the presence of ADP•AlF₄[−] (1 mM). ADP•AlF₄[−] was prepared by mixing in order the following molar ratio: 1 part Na-ADP (100 mM), 1 part AlCl₃ (1 M), and 5 parts NaF (500 mM). DHX36-Core-BeF₃ was crystallized by vapour diffusion at 21 °C using a reservoir consisting of 0.2 M potassium sodium tartrate (pH 7.4) and 20% (w/v) PEG 3350. ADP•BeF₃⁻ was prepared using the same protocol as above, except replacing AlCl₃ with BeSO₄. DHX36-Core-SeMet crystals were grown by vapour diffusion at 21 °C by combining 1.5 µl protein solution (5 g/l), 1 μl reservoir (50 mM sodium cacodylate (pH 7.1), 150 mM ammonium carbonate (pH 6.9), and 13.8% (v/v) 2-propanol), and 0.5 μ l microseed stock. The stock was made by crushing crystals of DHX36-core. All DHX36-core crystals were soaked in their respective reservoir solutions supplemented with 30% (v/v) glycerol before flash-freezing in liquid nitrogen. All DHX36-core crystals grew as rhombohedra to maximum dimensions of $500 \times 300 \times 300 \ \mu\text{m}^3$ in 1–2 weeks. DHX36-DSM complexed with DNA Myc was crystallized at 21 °C by hanging-drop vapour diffusion. Drops were prepared by combining complex solution (1.5 μl, 3 g/l), reservoir (1.0 μ l) and microseed stock (0.5 μ l). The reservoir consisted of 200 mM sodium malonate (pH 7.0) and 25% (w/v) PEG 3350. The seed stock was from crystals of unmethylated DHX36-DSM in complex with DNA Myc (grown in 45 mM MES-monohydrate (pH 5.7), 180 mM KCl, 29 mM MgCl $_2$, 4.5% (w/v) PEG 8000, 10 mM HEPES-NaOH (pH 7.5), and 3% (v/v) 2-propanol). Methylated DHX36-DSM–DNA $^{\textit{Myc}}$ crystals grew as plates to maximum dimensions of $500\times100\times5$ μm³ in 2–8 weeks. After growth, the reservoir solution was changed successively to 30% and 40% (w/v) PEG 3350 (other components unchanged) for a week each. Crystals, mounted on 90° bent loops (Mitegen), were flash-frozen in liquid nitrogen without further cryoprotection. Diffraction data were collected at 100 K in rotation mode with 0.9792 Å X-radiation at beamlines 5.0.1 and 5.0.2 of the Advanced Light Source (ALS), Lawrence Berkeley National Laboratory, and beamline 17-ID-B of the Advanced Photon Source (APS), Argonne National Laboratory. Data were indexed, integrated, and scaled using HKL2000³³ (Extended Data Tables 1, 2).

Structure determination and refinement. Data sets from five DHX36-core-SeMet crystals were scaled and merged together in HKL2000 (Extended Data Table 2). A heavy-atom substructure comprised of 18 selenium atoms was identified in this high-redundancy data set by HySS³⁴ implemented in PHENIX AutoSol³⁵. The resulting experimental SAD phases (mean overall figure of merit = 0.56) were density-modified with RESOLVE 36 to produce an electron density map in which manual model building using COOT³⁷ could begin (Extended Data Fig. 3). Iterative rounds of manual model building interspersed with rigid-body, simulated annealing, energy minimization, and individual isotropic B-factor refinement in PHENIX produced a near-complete model ($R_{\text{free}} = 35.2$) that could be placed (TFZ = 20.2) into the DHX36-core data set using PHASER³⁸. Further rounds of manual model building interspersed with refinement produced the current DHX36-core, DHX36core-AlF₄⁻, and DHX36-core-BeF₃⁻ models (Extended Data Table 1). Refined coordinates of the RecA1 domain from the 2.2 Å-resolution DHX36-core structure were used as a search model against the DHX36-DSM-DNA $^{\it Myc}$ data set, yielding $^{\rm 38}$ a solution with TFZ = 13.1. Subsequently, the RecA2 and C-terminal domains were successively placed (TFZs = 25.1 and 22.4, respectively). Rigid-body refinement followed by simulated annealing and restrained individual isotropic B-factor refinement was in turn followed by manual model building and further refinement to yield the current model of the DNA-protein complex (Extended Data Table 1). Coordinate precision estimates are from PHENIX. Structure figures were prepared with PyMol and Chimera 39,40 .

 $\textbf{Differential scanning calorimetry.} \ \ \textbf{Three DNA}^{Myc} \ \ \textbf{sequences (IDT, Extended)}$ Data Fig. 5a) at 0.1 mM concentration in 20 mM cacodylic acid-KOH (pH 7.2) and either 20 mM or 150 mM KCl were heated at 95 °C for 2.5 min, placed on ice for 10 min, and warmed to 21 °C over 20 min. The DNAs were analysed by size-exclusion chromatography (Superdex 75 Increase, GE Healthcare) in 20 mM cacodylic acid-KOH (pH 7.2) and 150 mM KCl (Extended Data Fig. 5b). For DSC, DNA samples prepared in 20 mM cacodylic acid-KOH (pH 7.2) and 20 mM KCl were degassed for 5-7 min before measurements (MicroCal VP-DSC differential scanning calorimeter). Thermograms were acquired between 20-105 °C at a scan rate of 0.5 °C min⁻¹ and at a constant pressure of 24 p.s.i. Three to five heating and cooling cycles were collected at least in duplicate for two independent preparations of each DNA sequence. Thermograms were highly reproducible (Extended Data Fig. 5c), and were analysed with Origin software (OriginLab). The reference 'buffer versus buffer' (20 mM cacodylic acid-KOH (pH 7.2) and 20 mM KCl) was subtracted from the sample data before curve-fitting (Levenberg-Marquardt nonlinear least-squares method) to determine $T_{\rm m}$ and ΔH .

Single-molecule FRET analyses. smFRET analyses of DHX36-DSM and site-directed mutants were performed as described^{22,28}.

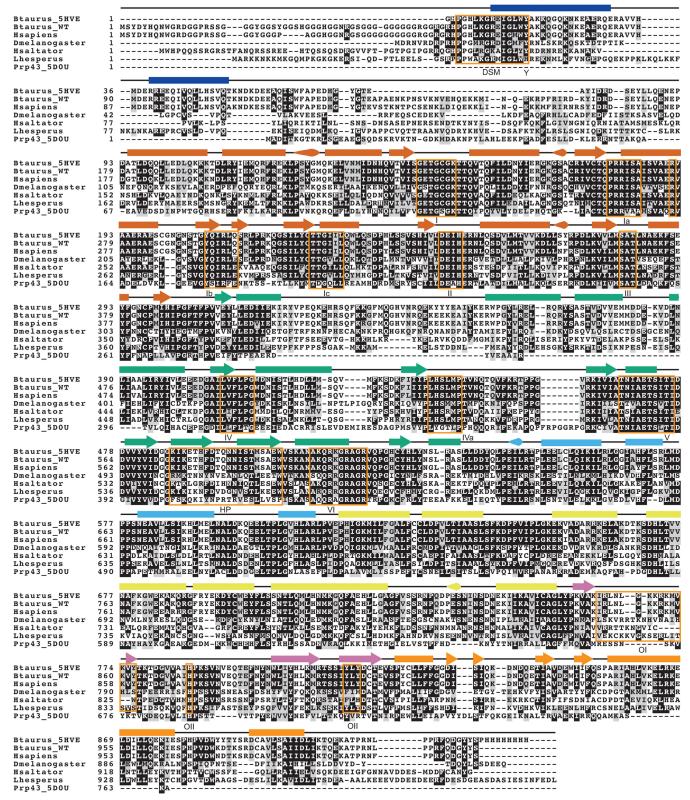
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Atomic coordinates and structure factors have been deposited in the RCSB Protein Data Bank (PDB) with accession numbers 5VHE for DHX36-DSM–DNA Myc , 5VHA for DHX36-core, 5VHC for DHX36-core-BeF $_3$, and 5VHD for DHX36-core-AlF $_4$.

- Andersen, K. R., Leksa, N. C. & Schwartz, T. U. Optimized *E. coli* expression strain LOBSTR eliminates common contaminants from His-tag purification. *Proteins* 81, 1857–1861 (2013).
- Studier, F. W. Protein production by auto-induction in high density shaking cultures. Protein Expr. Purif. 41, 207–234 (2005).
- 32. Rayment, I. et al. Three-dimensional structure of myosin subfragment-1: a molecular motor. *Science* **261**, 50–58 (1993).
- Otwinowski, Z. & Minor, W. Processing of diffraction data collected in oscillation mode. Methods Enzymol. 276, 307–326 (1997).
- Grosse-Kunstleve, R. W. & Adams, P. D. Substructure search procedures for macromolecular structures. *Acta Crystallogr. D Biol. Crystallogr.* 59, 1966–1973 (2003).
- Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D Biol. Crystallogr. 66, 213–221 (2010).
- Terwilliger, T. C. Maximum-likelihood density modification. Acta Crystallogr. D Biol. Crystallogr. 56, 965–972 (2000).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. 60, 2126–2132 (2004).
- McCoy, A. J. et al. Phaser crystallographic software. J. Appl. Crystallogr. 40, 658–674 (2007).
- 39. DeLano, W. L. The PyMOL Molecular Graphics System (DeLano Scientific, 2002).

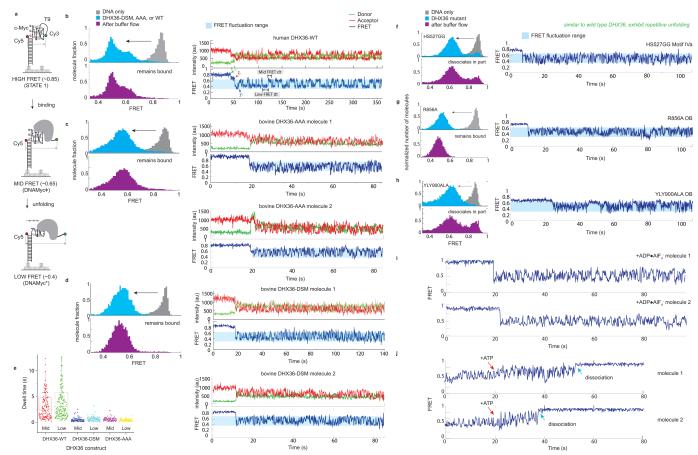


- Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612 (2004).
 Tauchert, M. J., Fourmann, J.-B., Christian, H., Lührmann, R. & Ficner, R. Structural and functional analysis of the RNA helicase Prp43 from the thermophilic eukaryote *Chaetomium thermophilum*. *Acta Crystallogr. F* 72, 112–120 (2016).
- Chalupníková, K. et al. Recruitment of the RNA helicase RHAU to stress granules via a unique RNA-binding domain. J. Biol. Chem. 283, 35186–35198 (2008).
 Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539 (2011).



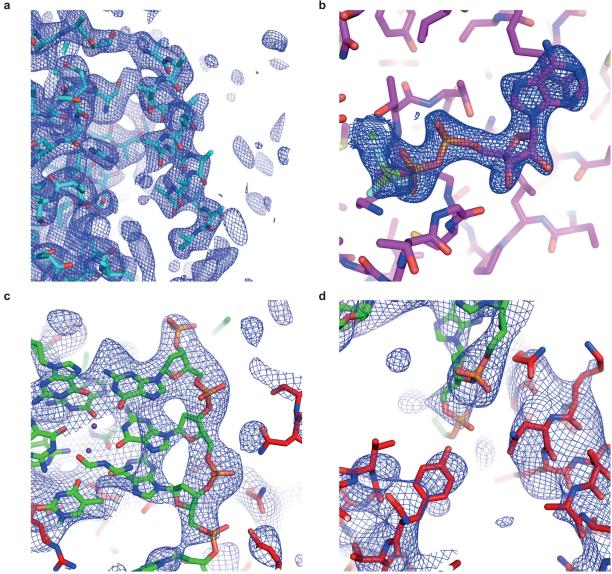
Extended Data Fig. 1 | Sequence alignment of DHX36 orthologues. The Bos taurus DHX36 construct used to solve the DHX36-DSM-DNA Myc co-crystal structure (PDB ID: 5VHE), wild-type Bos taurus DHX36, Homo sapiens DHX36, Drosophila melanogaster DHX36, Herpegnathos saltator DHX36, Latrodectus hesperus DHX36, and the Chaetomium thermophilum Prp43 crystallization construct 1 (PDB ID: 5D0U) are aligned with a 0.5 threshold for similarity (grey shading). The glycine-rich region is responsible for DHX36 recruitment to stress granules 2, but it is not

necessary for DHX36 binding or resolution of G-quadruplexes. Identical residues are shaded in black. Secondary structure from the DHX36-DSM-DNA Myc co-crystal structure is indicated above each alignment section, with arrow, rectangle and cone denoting α -helix, β -strand, and 3_{10} -helix, respectively. Secondary structure is colour-coded by domain or subdomain as in Fig. 1. Alignment was performed with Clustal Omega 43 and depicted using BoxShade (http://sourceforge.net/projects/boxshade/).



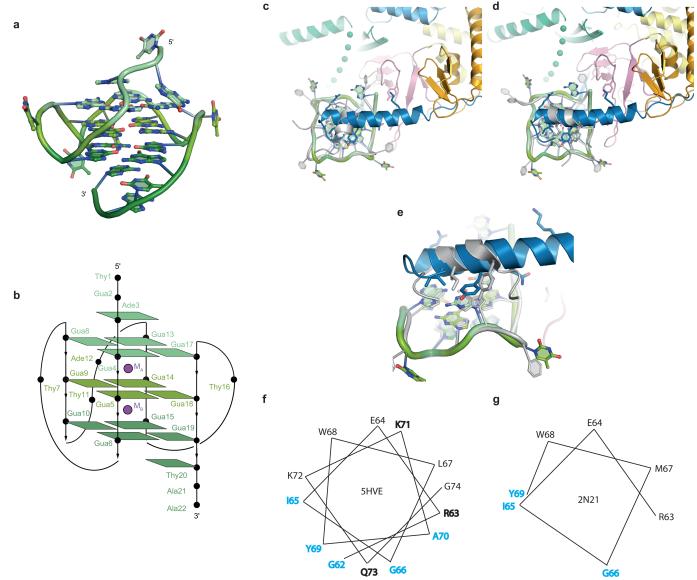
Extended Data Fig. 2 | Single-molecule FRET analysis of wild-type human DHX36 and bovine DHX36 constructs. a, Schematic of the smFRET assay^{22,28}. See Extended Data Fig. 8 for FRET state assignments. \mathbf{b} , Binding of wild-type human DHX36 (DHX36-WT) 22 to the G-quadruplex substrate, induces a shift from a high to medium and low FRET states (grey and cyan histograms, respectively). The shift is interpreted as the binding of DHX36 to the G-quadruplex substrate. Upon buffer flow, dissociation is not observed (purple histogram). Wildtype human DHX36 displays repetitive unfolding activity²², as indicated by the oscillation between medium and low FRET states after binding to the G-quadruplex substrate (blue trace). c, Binding of wild-type bovine DHX36 (incorporating a KKK192AAA mutation to prevent spontaneous proteolysis; DHX36-AAA) to the G-quadruplex substrate induces a shift from a high FRET state to medium and low FRET states (grey and cyan histograms, respectively). The shift is interpreted as the binding of DHX36 to the G-quadruplex substrate. Upon buffer flow, dissociation is not observed (purple histogram). Wild-type bovine DHX36 (DHX36-AAA) displays repetitive unfolding activity, as indicated by the oscillation between low and medium FRET states after binding to the G-quadruplex substrate (blue trace). FRET traces are shown for two molecules. d, Deletion of residues 111-159, mutation EEK435YYY, and mutation KDTK752AATA to generate DHX36-DSM does not impair G-quadruplex

binding or repetitive unfolding activity. FRET traces are shown for two molecules. e, Dwell time comparison between human DHX36-WT (grey bars), bovine wild-type DHX36 (DHX36-AAA, cyan bars) and bovine DHX36-DSM (orange bars). All three proteins show a comparable FRET range, and the two bovine constructs exhibit similar dwell times between the medium and low FRET states. Dwell times between the bovine constructs and the human construct are different, probably owing to interspecies differences. Each experiment was performed three times. Data are reported as box dot plots, with the data centre as the median \pm s.e. of 1,000 dwell times from 200 representative molecules. f-h, Mutation of motif IVa (hook loop) (f), the OB subdomain residue R856 (g), and OII does not result in impaired repetitive unfolding activity (h). However, partial dissociation following washing is observed with the motif IVa (f) and OII mutation (h). i, Pre-incubation of bovine DHX36-AAA with the non-hydrolysable ATP γ -phosphate hydrolysis transition state mimic ADP•AlF₄[−] does not affect repetitive unfolding activity on G-quadruplex substrates. j, Addition of ATP (red arrow) while DHX36-AAA is displaying repetitive unfolding activity on G-quadruplex substrates results in DHX36 dissociation (blue arrow) on the seconds timescale. Each experiment was repeated three times with highly similar results. Each measurement yields data from at least 10,000 molecules.

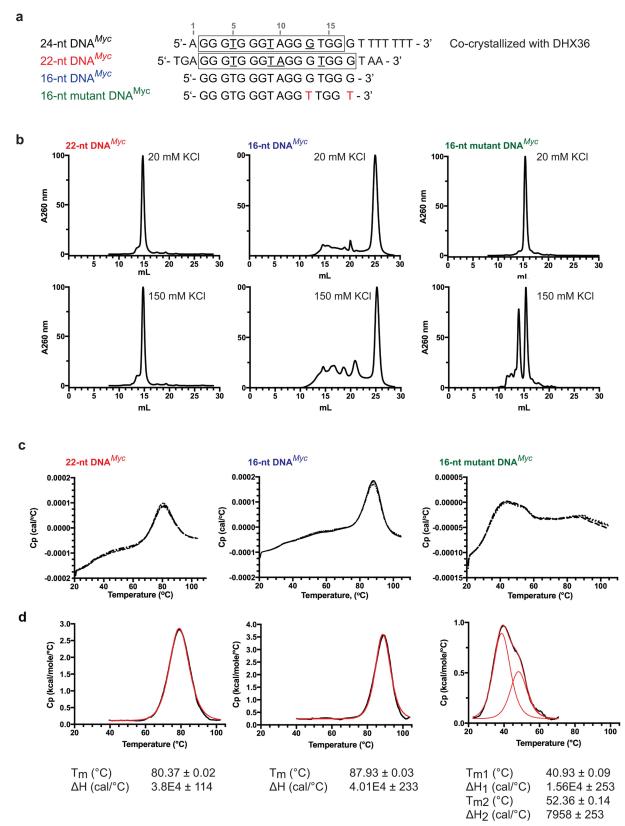


Extended Data Fig. 3 | Electron density maps superimposed on refined structures. a, Portion of the density-modified 3.1 Å resolution experimental SAD electron density map of selenomethionyl DHX36-core contoured at 1 s.d. above mean peak height, superimposed on a partially refined atomic model (see Methods). b, Portion of the 2.5 Å resolution simulated-annealing omit $2|F_0|-|F_c|$ electron density map of DHX36-core

in complex with ADP•BeF₃⁻ (PDB ID: 5VHC) contoured at 1.5. **c**, Portion of a simulated annealing-omit $2|F_0|-|F_c|$ electron density map of the DHX36-DSM-DNA^{Myc} complex corresponding to the G-quadruplex, contoured at 1 s.d. **d**, Portion of the electron density map (**c**) corresponding to the OI loop and the DSM helix (lower left and right, respectively). A portion of the DNA is in the upper centre.



align, the α -helix of the solution structure of the DSM-derived peptide is oriented approximately 90° with respect to the DSM α -helix from the DHX36-DSM–DNA Myc co-crystal structure. d, If arbitrarily rotated along the quadruplex four-fold axis, the DSM α -helices from both structures approximately align. e, Even with this rotation, the two structures differ in the DSM side chains presented to the DNA. f, g, Helical wheel representations of the DSM α -helices from the DHX36-DSM–DNA Myc co-crystal structure and the solution structure of the DSM-derived peptide bound to a G-quadruplex, respectively. Residues in cyan and bold make van der Waals contacts with the G-quadruplex face and hydrogen bond with the DNA backbone, respectively. Residue numbers correspond to the DHX36-DSM–DNA Myc co-crystal structure.

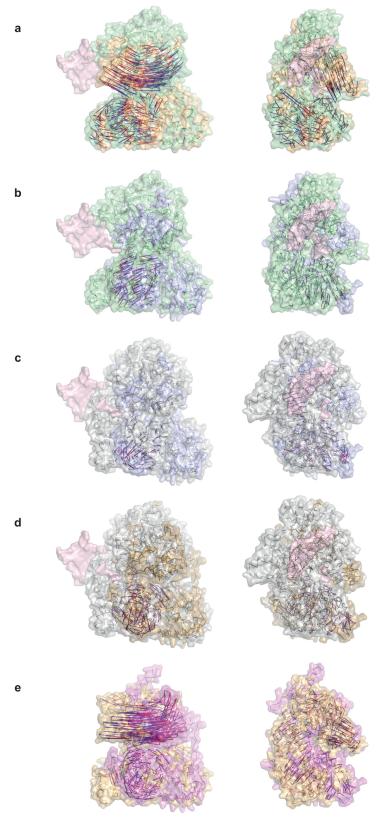


Extended Data Fig. 5 | See next page for caption.



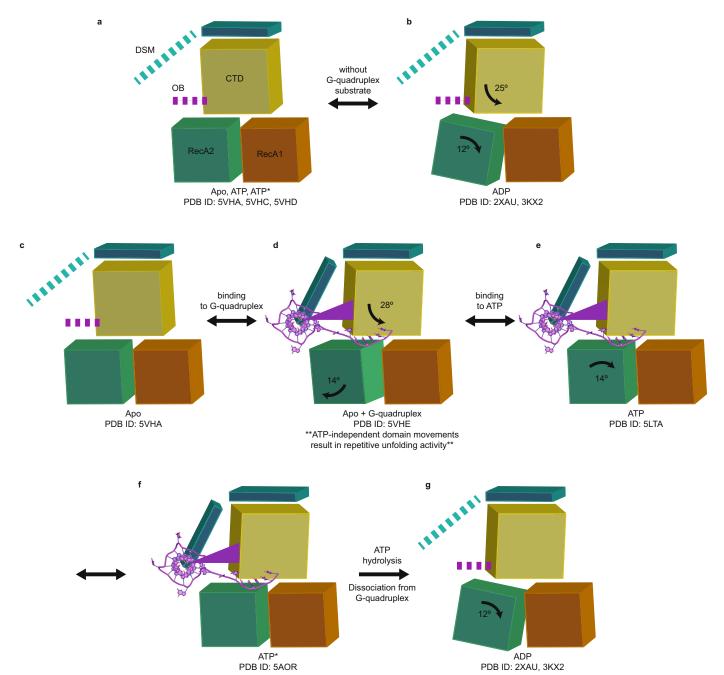
Extended Data Fig. 5 | Analysis of DNA^{Myc} conformers by differential scanning calorimetry (DSC). a, DNA constructs used in the analysis. DNA^{Myc}, DNA used for co-crystallization with DHX36-DSM (see Methods). Residues that form a three-tiered G-quadruplex in the complex and those that form propeller loops are boxed and underlined, respectively. 22-nt DNA^{Myc}, DNA used for solution NMR analysis²⁰. Residues that form a three-tiered G-quadruplex and those that form propeller loops in the free DNA are boxed and underlined, respectively. 16-nt DNA^{Myc}, DNA minimized to eliminate 5' and 3' single-stranded extensions to the G-quadruplex. 16-nt mutant DNA^{Myc}, variant of the former with two mutations (red) to enforce the three quartets observed

in the DHX36-DSM-DNA Myc co-crystal structure. **b**, Size-exclusion chromatograms (see Methods) of 22-nt DNA Myc , 16-nt DNA Myc and 16-nt mutant DNA Myc in the presence of either 150 mM or 20 mM KCl, demonstrating greater conformational homogeneity of the DNAs at lower KCl concentration. **c**, DSC thermograms (before buffer correction) for the three DNAs, in 20 mM KCl. Three independent experiments are plotted for each DNA. **d**, Triplicate nonlinear least-squares analyses of thermograms for the three DNAs. Black and red curves, buffer-corrected DSC data and curve-fits, respectively. $T_{\rm m}$ (melting temperature) and ΔH (enthalpy change) are reported as mean \pm s.d. Each experiment was repeated three times with two sets of identical DNA preparations.



Extended Data Fig. 6 | Alignments of the structures of DHX36, MLE, and Prp43. RecA1 domains were superimposed. Vectors from red to blue denote $C\alpha$ displacement between identical or structurally homologous residues. a, Superposition of DHX36-DSM-DNA^{Myc} and unliganded DHX36-core (5VHA) structures (green and orange, respectively). DNA^{Myc} is pink. b, Superposition of DHX36-DSM-DNA^{Myc} (green) and Prp43 (ref. ¹⁶) bound to rU₁₆ and ADP \bullet BeF₃ (5LTA; blue; ground'). DNA^{Myc} from the DHX36-DSM-DNA^{Myc} structure is pink. c, Superposition of

Prp43 bound to rU₈ and ADP \bullet BeF₃ ⁻ (5LTA; blue; 'ground') to MLE¹⁵ bound to rU15 and ADP \bullet AlF₄ ⁻ (5AOR; silver; 'transition'). DNA^{Myc} from the DHX36-DSM-DNA^{Myc} structure is pink. **d**, Superposition of MLE bound to rU₁₅ and ADP \bullet AlF₄ ⁻ (5AOR; silver; 'transition') and Prp43 bound^{13,14} to ADP (3KX2/2XAU; gold; 'post-hydrolysis'). DNA^{Myc} from the DHX36-DSM-DNA^{Myc} structure is pink. **e**, Superposition of Prp43 bound to ADP (3KX2/2XAU; gold; 'post-hydrolysis') to unliganded DHX36-core (5VHA; magenta; 'apo').



Extended Data Fig. 7 | Model of the mechanochemical cycle of the DEAH/RHA helicase DHX36. The domain motions are based on the superpositions in Extended Data Fig. 6. The orange, green, yellow, and blue blocks represent the RecA1 domain, RecA2 domain, C-terminal domain, and N-terminal extension, respectively. The purple wedge represents the OB domain. Bold dotted lines represent likely intrinsically disordered protein motifs that fold upon G-quadruplex binding.

a, b, In the absence of a G-quadruplex nucleic acid substrate, DHX36 cycles between an apo (or structurally indistinguishable ATP-bound) state and a post-hydrolysis state. c, d, DHX36 binds the G-quadruplex substrate and pulls on it in the 3'-direction through concerted and opposite rotations of the RecA2 and C-terminal domains. Oscillation of the RecA2 and C-terminal domains is likely to be responsible for the

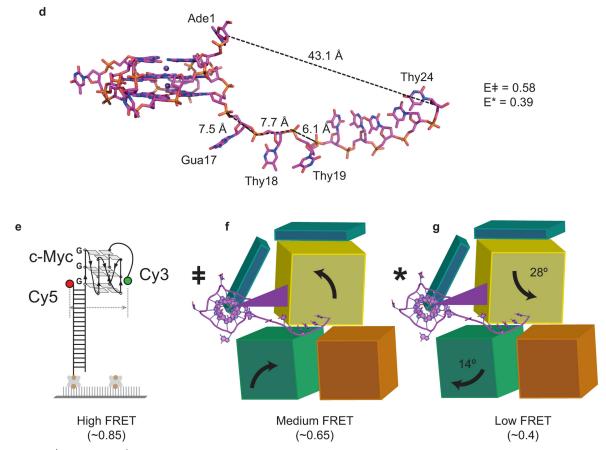
ATP-independent repetitive unfolding activity detected by smFRET 22 (Extended Data Fig. 2 and Fig. 4). **d**, **e**, Binding of ATP induces domain closure. **f**, **g**, ATP hydrolysis yields a post-hydrolysis state that is incompatible with nucleic acid binding. ADP dissociates, and DHX36 is reset back to its apo state (**c**). In addition to the rearrangement of motif Va^{17} , ATP hydrolysis is stimulated by nucleic acid binding, probably because nucleic acid binding results in the opening of the helicase core. Diffusion into the NTP binding pocket is thus increased. The model in **e** is based on the superposition in Extended Data Fig. 6b. The model in **f** is based on the superposition in Extended Data Fig. 6c. The model in **g** is based on the superposition in Extended Data Fig. 6c. The model in **g** is based on the superposition in Extended Data Fig. 6d.

a b

C

DNAMyc + 5' - A GGG T GGG TA GGG T GGG TTTTTTT - 3'

DNAMyc* 5' - A GGG T GGG T AGG GTTTTTTT - 3'



Extended Data Fig. 8 \mid See next page for caption.

Extended Data Fig. 8 | Comparison of canonical and reorganized DNA Myc G-quadruplex. DNA Myc ‡ denotes the canonical DNA Myc structure 20,24 whereas DNA Myc * represents the reorganized DNA Myc found in the DHX36-DSM-DNA Myc co-crystal structure. **a**, Structure of the DNA Myc ‡ top G-quartet (PDB ID: 2N21). **b**, Structure of the DNA Myc * top G-quartet. **c**, Primary sequence alignment of the canonical and reorganized DNA Myc G-quadruplex. Bold residues participate in formation of a quartet. **d**, The structure of DNA Myc G-quadruplex found in our co-crystal structure, represented here by DNA Myc *. Distances between A1 and T24 as well as G16 and G17, G17 and T18, and T18 and T19 are indicated. Theoretical FRET efficiencies (E) for DNA Myc ‡ and DNA Myc * were calculated using $E = 1/[1 + (r/R_0)^6]$ where $R_0 = 53$ Å for the Cy3-Cy5 pair and r is the distance between Cy3 and Cy5. Since smFRET experiments were performed with a DNA Myc G-quadruplex containing a 3' ssDNA extension of nine thymines, we added the distance between

two thymines to the theoretical FRET efficiency model assuming an average internucleotide distance of 7.1 Å. As the difference between the hypothetical DNA^{Myc}‡ previously solved by NMR and DNA^{Myc}* found in our co-crystal structure is one nucleotide, we modelled r‡ and r* as 50.2 Å and 57.3 Å, respectively. From these parameters, we obtained predicted FRET efficiencies of 0.58 and 0.39 for DNA^{Myc}‡ and DNA^{Myc}*, respectively. These predicted FRET efficiencies closely match the experimental oscillating FRET efficiencies of ~0.6 and ~0.4. e, The high FRET state of ~0.85 is observed before DHX36 binding to the DNA^{Myc} G-quadruplex. f, DHX36 initially binds to DNA^{Myc}‡ (FRET ~0.6). g, Probably owing to ATP-independent C-terminal domain rotations also observed with Prp43p, the DNA^{Myc}, G-quadruplex is partially unwound to DNA^{Myc}* (~0.4). DHX36 then oscillates between DNA^{Myc}* and DNA^{Myc}‡ in an ATP-independent repetitive unfolding activity.



Extended Data Table 1 \mid Data collection and refinement statistics

	DHX36-DSM- DNA ^{Myc}	DHX36-Core (PDB: 5VHA)	DHX36-Core-BeF ₃ -(PDB: 5VHC)	DHX36-Core-AlF ₄ (PDB: 5VHD)
	(PDB: 5VHE)			
Data collection				
Space group	$P2_1 2_1 2_1$	$P2_1$	$P2_1$	$P2_1$
Cell dimensions				
a, b, c (Å)	72.5, 79.3, 212.1	61.3, 109.2, 62.4	61.9, 111.5, 63.0	62.0, 112.5, 63.2
α, β, γ (°)	90, 90, 90	90, 112.7, 90	90, 110.7, 90	90, 110.3, 90
Resolution (Å)	39.6-3.8 (3.9-3.8) ^a	39.6-2.2 (2.3-2.2)	37.8-2.5 (2.6-2.5)	46.7-2.6 (2.7-2.6)
R_{merge} (%)	27.5 (168)	8.66 (46.7)	7.52 (25.9)	15.0 (80.0)
< <i>I</i> >/<σ(<i>I</i>)>	7.8 (1.3)	18.3 (3.0)	16.9 (3.6)	9.9 (1.2)
$CC_{1/2}$	0.994 (0.486)	0.997 (0.883)	0.996 (0.949)	0.994 (0.646)
Completeness (%)	99.2 (94.0)	99.8 (98.1)	91.5 (87.8)	98.8 (98.2)
Redundancy	9.6 (7.6)	6.5 (6.2)	3.5 (3.4)	5.1 (3.6)
Refinement				
Resolution (Å)	39.6-3.8 (3.9-3.8)	39.6-2.2 (2.3-2.2)	37.8-2.5 (2.6-2.5)	46.7-2.6 (2.6-2.6)
No. reflections	12606 (1153)	37004 (3636)	25665 (2446)	26260 (2612)
$R_{ m work}$ / $R_{ m free}$ (%)	23.8/28.0	17.7/21.7	19.4/23.3	17.3/21.3
No. atoms	7202	6477	6318	6266
Protein	6697	6286	6189	6126
DNA	503	0	0	0
Ligand/ion	2	0	32	32
Water	0	180	97	108
Mean <i>B</i> -factors ($Å^2$)	120.9	53.6	60.9	50.1
Protein	118.4	53.8	61.8	50.0
DNA	155.2	N/A	N/A	N/A
Ligand/ion	117.5	N/A	47.2	78.8
Water	N/A	47.3	53.3	45.5
R.m.s. deviations				
Bond lengths (Å)	0.002	0.004	0.002	0.002
Bond angles (°)	0.50	0.75	0.52	0.56
Ramachandran analysis (%)				
Favored	91.2	96.3	97.2	96.8
Allowed	7.5	3.2	2.8	3.2
Disallowed	1.3	0.5	0.0	0.0
Mean coordinate precision (Å)	0.47	0.30	0.32	0.21

^aValues in parentheses are for highest-resolution shell. One crystal was used for each of the four data sets.

Extended Data Table 2 \mid Data collection statistics for DHX36-core-SeMet crystals

	DHX36-	DHX36-	DHX36-	DHX36-	DHX36-	DHX36-
	Core-	Core-	Core-	Core-	Core-	Core-SeMet
	SeMet 1	SeMet 2	SeMet 3	SeMet 4	SeMet 5	<1,2,3,4,5>
Data collection						_
Space group	$P2_1$	$P2_1$	$P2_1$	$P2_1$	$P2_1$	$P2_1$
Cell dimensions						
a, b, c (Å)	62.6, 115.7,	62.5, 113.6,	62.5, 113.4,	62.4, 113.5,	62.6, 114.7,	62.6, 114.7,
	64.0	63.8	63.8	64.0	63.9	63.9
α, β, γ (°)	90, 107.3, 90	90, 108.0,	90, 108.0,	90, 107.8, 90	90, 107.8, 90	90, 107.8, 90
••••		90	90			
Resolution (Å)	46.7-3.2	46.6-3.3	46.6-3.0	46.6-2.9	46.7-3.1	46.7-3.1
	$(3.3-3.2)^a$	(3.4-3.3)	(3.1-3.0)	(3.0-2.9)	(3.2-3.1)	(3.2-3.1)
R_{merge} (%)	17.4 (87.6)	17.4 (94.1)	14.5 (84.8)	14.8 (74.8)	14.4 (81.5)	22.4 (102)
< <i>I</i> >/< $\sigma(I)$ >	10.5 (1.3)	10.8 (1.5)	11.6 (1.2)	11.0 (1.4)	12.1 (1.4)	22.7 (2.1)
$CC_{1/2}$	0.995	0.996	0.997	0.997	0.997	0.999
	(0.643)	(0.646)	(0.743)	(0.745)	(0.798)	(0.753)
Completeness (%)	97.5 (80.5)	98.6 (88.6)	92.5 (60.8)	90.6 (51.2)	97.9 (12.1)	98.9 (90.1)
Redundancy	7.0 (5.0)	7.0 (4.9)	6.7 (4.8)	6.7 (4.3)	7.0 (5.0)	31.1 (15.5)

^aValues in parentheses are for highest-resolution shell.

One crystal was used for each of the first five data sets. The last data set resulted from merging all five data sets.



Structural basis for regulation of human acetyl-CoA carboxylase

Moritz Hunkeler^{1,4,5}*, Anna Hagmann^{1,5}, Edward Stuttfeld¹, Mohamed Chami^{1,2}, Yakir Guri¹, Henning Stahlberg^{1,3} & Timm Maier¹*

Acetyl-CoA carboxylase catalyses the ATP-dependent carboxylation of acetyl-CoA, a rate-limiting step in fatty acid biosynthesis^{1,2}. Eukaryotic acetyl-CoA carboxylases are large, homodimeric multienzymes. Human acetyl-CoA carboxylase occurs in two isoforms: the metabolic, cytosolic ACC1, and ACC2, which is anchored to the outer mitochondrial membrane and controls fatty acid β -oxidation^{1,3}. ACC1 is regulated by a complex interplay of phosphorylation, binding of allosteric regulators and protein-protein interactions, which is further linked to filament formation^{1,4-8}. These filaments were discovered in vitro and in vivo 50 years ago^{7,9,10}, but the structural basis of ACC1 polymerization and regulation remains unknown. Here, we identify distinct activated and inhibited ACC1 filament forms. We obtained cryoelectron microscopy structures of an activated filament that is allosterically induced by citrate (ACC-citrate), and an inactivated filament form that results from binding of the BRCT domains of the breast cancer type 1 susceptibility protein (BRCA1). While nonpolymeric ACC1 is highly dynamic, filament formation locks ACC1 into different catalytically competent or incompetent conformational states. This unique mechanism of enzyme regulation via large-scale conformational changes observed in ACC1 has potential uses in engineering of switchable biosynthetic systems. Dissecting the regulation of acetyl-CoA carboxylase opens new paths towards counteracting upregulation of fatty acid biosynthesis in disease.

Eukaryotic acetyl-CoA carboxylases comprise biotin carboxylase (BC), biotin carboxyl carrier protein (BCCP) and carboxyl transferase (CT) domains, as well as an interaction domain (BT) and a non-catalytic central domain region (CD), which together bridge the BC and CT domains¹ (Fig. 1a). The CD comprises four domains, the N-terminal CD_N, the linking CD_L, and the tandem C-terminal CD_{C1} and CD_{C2}¹¹. Acetyl-CoA carboxylation is a two-step reaction: first, a BCCP-linked biotin moiety is carboxylated with ATP consumption by the BC domain; second, the resulting carboxybiotin is shuttled to the CT domain and the carboxy group is transferred onto acetyl-CoA. In fungal acetyl-CoA carboxylase, site-specific phosphorylation in the CD regulates activity by controlling the transition between an inactive, open state and an active, closed form, which is characterized by dimerization of BC domains^{11,12}. Human ACC1 (Fig. 1b) (hereafter referred to as ACC) is inactivated by phosphorylation at Ser80, Ser1201 and Ser1216 by AMP-activated protein kinase (AMPK) and at Ser78 and Ser1201 by cAMP-dependent protein kinase (PKA); the Ser80 and Ser1201 sites have the largest impact on activity⁴. ACC is further inhibited by its product malonyl-CoA and the fatty acid derivative palmitoyl-CoA⁵. The allosteric activator citrate induces polymerization of ACC into unbranched filaments of up to 1 μm in length⁷; these filaments are the most active form of ACC^{6,7}. Additionally, the tumour suppressor BRCA1 has been hypothesized to regulate ACC¹³. BRCA1 binds to ACC through its C-terminal tandem BRCT domains¹⁴ (Fig. 1a), which recognize phosphorylated Ser1263 in the CD of ACC¹⁵. Ser1263 is phosphorylated in a cell cycle-dependent manner, presumably by a cyclin-dependent kinase^{13,15}. Binding of BRCA1 prevents dephosphorylation of Ser80 and thus inhibits activation of ACC¹³. Mutations in the BRCT domains abolish BRCA1 binding to ACC, resulting in elevated lipogenesis, which is a prerequisite for cancer cell growth^{13,14}.

We expressed ACC in insect cells; mass spectrometry confirmed that the expressed protein was phosphorylated on Ser80 (76%) and Ser1263 (94%). Dephosphorylated protein was obtained by λ -phosphatase treatment, and was five times more active than the phosphorylated protein (Fig. 1c). Addition of citrate to dephosphorylated ACC induces formation of ACC–citrate filaments (Fig. 1d). Addition of palmitoyl-CoA to preformed ACC–citrate filaments at tenfold molar excess, which is sufficient to inhibit ACC (Fig. 1e), induces a transition to another, apparently related filament form (ACC–citrate^{palm}) (Fig. 1d). BRCT binding to phosphorylated ACC yields an ACC filament (ACC–BRCT) that has a distinct architecture from ACC–citrate or ACC–citrate^{palm} filaments (Fig. 1d, Extended Data Fig. 1).

ACC filaments are highly flexible and form a clustered meshwork on electron microscopy grids (Extended Data Fig. 2a). Extensive screening yielded suitable samples for cryo-electron microscopy (cryo-EM) studies of ACC-citrate and ACC-BRCT filaments. Helical symmetry was evident from the raw images; however, helical processing was not applied owing to the large repeating unit and the pronounced curvature of the filaments. Single-particle analysis yielded reconstructions at resolutions between 4.6 Å and 5.9 Å, as judged by Fourier shell correlation (FSC, 0.143 threshold criterion) for different filaments and filament regions (Extended Data Table 1, Extended Data Figs. 2, 3). Map quality of subregions was improved using local post-map symmetry averaging (Extended Data Fig. 4). Except for the peripheral BRCT and BC regions in the ACC-BRCT filament, all secondary structure elements were clearly resolved. Crystal structures of individual domains were fitted into cryo-EM maps. Dimeric BRCT domains were placed unambiguously on the basis of their interaction with phosphorylated (p) Ser1263, which is located on a flexible loop (amino acids 1257–1283) in the CD_{C1} domain of ACC, despite the lower local resolution (Extended

Upon addition of citrate, dephosphorylated ACC assembles into filaments of 0.5–1 μm in length with a helical twist of approximately 120° and a rise of 154 Å (Fig. 2a, b, Supplementary Video 1). In the cryo-EM maps, all domains, including the flexibly tethered carrier protein, are resolved. The filament assembles by lateral stacking of ACC dimers, which resemble the triangular, closed conformation previously observed for yeast ACC (Fig. 2c): Two BC domains form a dimer, which has been recognized as a prerequisite for BC activity 12 . Citrate-induced filament formation therefore locks ACC into an active conformation, which is not highly populated for non-filamentous ACC in conditions lacking citrate (Extended Data Fig. 5a) or for residual non-polymerized ACC in citrate-containing conditions (Extended Data Fig. 5b, 5c).

In the two-step ACC reaction, the biotin moiety on BCCP is carboxylated by the BC domain and then shuttled to the CT domain. Notably, in the cryo-EM map, BCCP is unambiguously located at the carboxyl

¹Biozentrum, University of Basel, Basel, Switzerland. ²BioEM Lab, Biozentrum, University of Basel, Basel, Switzerland. ³Center for Cellular Imaging and NanoAnalytics, Biozentrum, University of Basel, Basel, Switzerland. ⁴Present address: Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. ⁵These authors contributed equally: Moritz Hunkeler, Anna Hagmann. *e-mail: moritz hunkeler@dfci.harvard.edu; timm.maier@unibas.ch

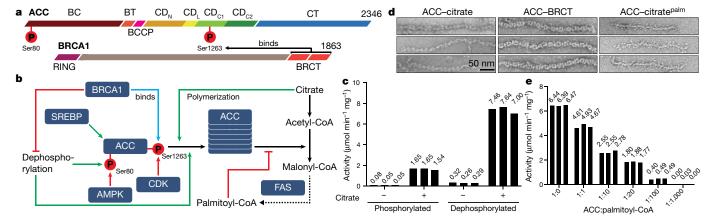


Fig. 1 | Domain organization and regulation of ACC. a, Domain organization of ACC and BRCA1; the colour scheme is used throughout the manuscript. Selected ACC phosphosites are indicated. The BRCT domains of BRCA1 have been reported to interact with ACC via pSer1263. b, Overview of ACC regulation. Activation and polymerization are shown in green, inhibition in red and BRCA1 binding in blue. Sterol regulatory

element-binding protein (SREBP) controls ACC expression; other effects are at the protein level. **c**, Specific activity of phosphorylated and dephosphorylated ACC in presence and absence of citrate. **d**, Negative stain electron micrographs of three types of ACC filaments. Scale bar, 50 nm. **e**, Activity of ACC in presence of citrate and palmitoyl-CoA. Three individual measurements are shown for each condition in **c** and **e**.

transferase active site with ordered flanking BCCP linkers (Fig. 2d), strongly indicating that the CT domain acts as a docking platform for BCCP in the resting state of ACC–citrate filaments. Additional density is present for the biotin moiety linked to BCCP (Extended Data Fig. 5d–f). The biotin carboxylase active site is located 80 Å from the carboxyl transferase active site. A simple 120° rotation of BCCP around hinges in the connecting linkers is sufficient to switch its position between the two active sites (Fig. 2d).

Inter-dimer interactions in ACC–citrate filaments are mediated by the CD. The CD $_N$ domain of ACC molecule A contacts CD $_L$ –CD $_{C1}$ of molecule A+1, and vice versa (Fig. 2a, Extended Data Fig. 6a). This interface is based on the docking of a loop between helices N $\alpha 4$ and N $\alpha 5$ of CD $_N$ into a cradle formed by strand $\beta 1$ of the β -sheet of CD $_{C1}$ and helices L $\alpha 2$ and L $\alpha 4$ of CD $_L$ (Extended Data Fig. 6a–c). The patches involved in this interface are highly conserved amongst metazoans, consistent with previous observations of polymers of avian, murine and bovine ACC 7,16 . However, they are less conserved for fungal ACCs (Extended Data Fig. 6d). Superimposing the structure of Saccharomyces cerevisiae ACC (ScACC, PDB ID: 5CSL) indeed suggests that the yeast enzyme is incompatible with the observed mode of

filament formation (Extended Data Fig. 6b, c). It cannot be excluded, however, that flexibility of involved loop regions or slight variations in helical symmetry might allow ACC polymerization in fungi. To our knowledge, no polymeric forms of purified fungal ACCs have been reported. Two recent studies visualized fungal ACC in vivo in elongated foci using wide field or confocal microscopy^{17,18}, however, they did not provide further evidence for direct polymerization of ACC.

The activated ACC-citrate filament selects for a dimeric BC arrangement, although the BC domains are not directly involved in filament assembly. Conformational restriction of the CD is apparently sufficient to stabilize BC domain dimerization. Nevertheless, disturbed BC domain dimer interfaces, for example, in response to Ser80 phosphorylation, as previously proposed¹², could still be compatible with this filament form (Extended Data Fig. 1).

Addition of a tenfold molar excess of the feedback inhibitor palmitoyl-CoA reduces activity of citrate-activated ACC by approximately 60% (Fig. 1e). It also disturbs the architecture of preformed ACC-citrate filaments, resulting in another filament form, which we have termed ACC-citrate^{palm} (Fig. 1d). ACC-citrate^{palm} forms within minutes from ACC-citrate and is stable for several hours.

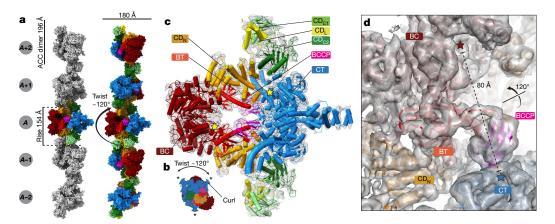


Fig. 2 | **ACC-citrate filament structure. a**, ACC-citrate filaments (surface representation) assemble from closed ACC dimers. Rise, twist, width and ACC dimer extent are indicated. Left, filament with one dimer in domain colours; right, filament coloured by domains. **b**, Top view of ACC-citrate filament. Helical curls are marked by asterisks. **c**, Closed ACC dimer in ACC-citrate filaments shown with cryo-EM map at contour level 0.0172. Yellow stars mark active sites. Domains of one protomer are labelled.

d, The BCCP domain is positioned at the carboxyl transferase active site. One ACC-citrate dimer (cartoon representation) is shown together with the cryo-EM map at contour level 0.0158. Carboxyl transferase and biotin carboxylase active sites are marked by blue and red stars; the distance between the sites is indicated. BCCP linkers are shown as bold tubes. The 120° rotation moving the BCCP to the BC domain is indicated.

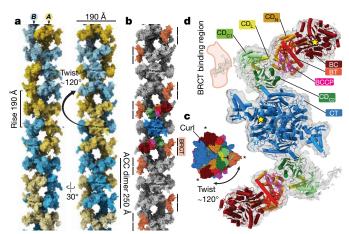


Fig. 3 | **ACC-BRCT filament structure. a**, The two strands in ACC-BRCT filaments (surface representation) are shown in blue and yellow, respectively. Rise and twist for one strand, width and ACC dimer extent are indicated. **b**, The BRCT dimer (coral, indicated) is positioned at the periphery of the filament. The filament is shown in grey with one ACC dimer coloured by domains. **c**, Filament top view illustrating the triangular shape and central positioning of the α -helical curl. Helical twist is indicated. **d**, Inverse Z-shaped ACC dimer shown with cryo-EM map. Yellow stars mark active sites. Domains of one protomer are indicated. Map contour level (0.007) was chosen to reveal the density of all domains.

ACC-citrate^{palm} filaments begin to dissociate only at considerably higher palmitoyl-CoA concentrations, possibly owing to palmitoyl-CoA acting as a detergent (Extended Data Fig. 7a). In ACC-citrate^{palm}, the helical backbone is thinned, and globular satellites flank the sides of filaments. We assume that the two filament forms are related and that the inter-protomeric backbone of ACC-citrate is also present in ACC-citrate^{palm}. The most likely candidate for the globular satellites is the BC domain, which has been observed in dimeric and monomeric forms as part of fungal ACC¹⁹ or excised from human ACC2²⁰ and is not part of the filament spine. We hypothesize that binding of palmitoyl-CoA leads to conformational changes, possibly in the CD or the BC domain dimer interface, resulting in BC dimer destabilization (Extended Data Fig. 7b) and, therefore, reduced catalytic activity.

Upon binding of the BRCT domains, phosphorylated ACC forms twostranded filaments with a triangular cross-section and, for each strand, a helical twist of approximately 120° and a rise of 190 Å (Fig. 3a-c and Supplementary Video 2). Each strand assembles from open Z-shaped ACC dimers that are aligned along the filament axis (Fig. 3d). Dimeric BRCT domains laterally decorate the filament (Fig. 3b) and interlink adjacent ACC dimers. The filament consists of protein-dense nodes interconnected by arm-like protrusions (Extended Data Fig. 7c-e). Owing to the extended Z-shape, each dimer contributes to three consecutive nodes, and each node is composed of the two CT domains of molecule A, one copy each of the CD_{C1}, CD_L, CD_N, BT, BCCP and BC domains of molecules B and B-1, as well as a dimer of BRCT domains interlinking the CD_{C1} domains of molecules *B* and *B*−1 (Fig. 4a, Extended Data Fig. 7f, g). The connecting arms are formed by CD_{C2} domains (Extended Data Fig. 7e), which contact the CD_N and CD_{C1} domains of the same protomer in one node and the CT domain in the preceding node with interface areas of 430 $Å^2$, 870 $Å^2$ and 450 $Å^2$, respectively.

The inter-strand interactions are formed in the centre of the nodes, where four-helix bundles of the α -helical CD_N domains from molecules B and B-1 form an interface of approximately 900 Ų with the C-terminal α -helical curl and the preceding α -helical extensions of the CT domain of molecule A (Fig. 4a). The helical curl had previously been linked only to stabilization of the CT domain dimer and compound binding²¹, but apparently has a central role in organizing the ACC–BRCT filament. The C-terminal 85 amino acids of the CT domain are indeed highly conserved between human and other mammalian ACCs with 52–75% pairwise sequence identity, whereas the

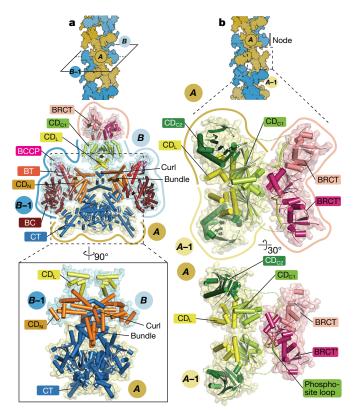


Fig. 4 | Inter- and intra-strand interactions in ACC–BRCT filaments. a, Top, ACC–BRCT filament with protomers from two strands labelled and cross-section plane indicated. Middle, cross-section reveals binding of the $CD_{\rm N}$ four-helix bundle to the α -helical extension and the curl as the main inter-strand interaction. One instance of each domain is labelled, coloured outlines indicate protomer connectivity. Bottom, enlarged view (rotated) of interaction area. b, Top, ACC–BRCT filament with protomers from one strand labelled. Middle and bottom, enlarged view of the $CD_{\rm Cl}$ –BRCT interaction. Each protomer of dimeric BRCTs binds to one phosphosite loop. Outlines show connectivity, dashed lines indicate non-modelled residues. The phosphosite loop is shown in bold. Extended Data Fig. 7g shows a further enlarged view.

sequence identity to the C terminus of the CT domain in yeast ACC is only $13\%^{21}$. The four-helix bundle of the CD_N domain is involved both in the inter-strand interactions in ACC–BRCT filaments and in the inter-dimer interfaces of ACC–citrate. The BC domain in ACC–BRCT is monomeric, therefore ACC–BRCT filaments represent an inhibited form of ACC^{11,12,19}. Within a node (Fig. 4a), the distance between the active sites of the BC domains of molecules *B* and *B*–1 is 143 Å and the distance to the active site in the CT domain of molecule *A* is only 70 Å, less than the corresponding distance in the active ACC–citrate filament. However, even if the BC domains were in a catalytically competent state despite being monomeric, the BCCP domains would be sterically unable to reach any of the active sites.

The inverse Z-shape of ACC dimers in ACC–BRCT is an intermediate between the open, inactive conformation of *Chaetomium thermophilum* and yeast ACC²² and the closed, active form of non-polymeric yeast ACC or the human ACC–citrate filaments (Extended Data Fig. 8a). The distinct conformations of human ACC dimers arise from bending of hinges in the CD (Extended Data Fig. 8b–d). In ACC–BRCT, the BC domain is located so that its B-domain cap resides in the C-shaped CD_N domain (Extended Data Fig. 8e) with increased conformational variability, based on map quality. Notably, the positions of the BT and BCCP domains relative to CD_N are largely conserved between ACC–citrate and ACC–BRCT (Extended Data Fig. 8f). However, the BCCP domain in ACC–BRCT, analogous to the BC domain, is more flexible and not docked to the carboxyl transferase active site as in ACC–citrate (Fig. 2d, Extended Data Fig. 8g).

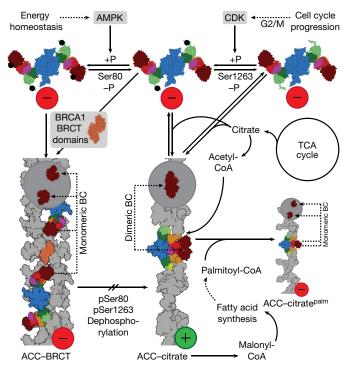


Fig. 5 | ACC polymerization integrates regulatory signals. AMPK phosphorylates Ser80 of dimeric ACC in response to the cellular energy state, whereas Ser1263 is phosphorylated in a cell cycle-dependent manner. Upon binding of citrate, ACC with unphosphorylated Ser80 forms activated ACC-citrate filaments with dimeric BC domains. ACC yields malonyl-CoA for fatty acid biosynthesis. The feedback inhibitor palmitoyl-CoA modifies ACC-citrate filaments, potentially by disrupting BC domain dimerization. Binding of BRCT domains of BRCA1 at pSer1263 results in formation of inactive ACC-BRCT filaments with monomeric BC domains. Grey circles show enlarged views of BC organization. Binding of full-length BRCA1 inhibits dephosphorylation of pSer80 and pSer1263.

The intra-strand connections in ACC–BRCT are mediated by CD_{C1} domains of successive ACC molecules. These domains form only a minimal contact area of 100 Ų. ACC–BRCT filaments are not observed in the absence of the BRCT domains (Extended Data Fig. 1), indicating that the CD_{C1} contact is not sufficient to independently establish stable polymerization (Fig. 4b). The CD_{C1} domain comprises the phosphosite loop (amino acids 1257–1283), which is disordered in the crystal structure of the BT–CD region of ACC^{11} and contains the regulatory pSer1263. The phosphosite loop provides the binding site for BRCT, as previously demonstrated in the structure of a monophosphorylated peptide mimic (1258-DSPPQ-pS-PTFPEAGH-1271) bound to monomeric BRCT¹⁵.

In ACC-BRCT, however, a dimer of BRCT domains binds to the (partially disordered) phosphosite loops of the CD_{C1} domains of successive ACC dimers in one strand. The BRCT domains interact with an interface area of approximately 800 Å² in the BRCT dimer, which is held in place mostly by interactions with the protruding phosphosite loops (Fig. 4b, Extended Data Fig. 7f, g), and only marginally contacts the ACC filament core (interface area approximately 100 Å²). An equivalent mode of BRCT dimerization has been observed for interactions of BRCT with peptides of the abraxas protein: Monophosphorylated abraxas peptides interact with monomeric BRCT, whereas peptides that contain a second pSer preceding the first phosphosite form dimeric BRCT-phosphopeptide complexes²³. In ACC, Ser1259 precedes the canonical pSer1263 recognition site of BRCT. Phosphorylation of the Ser1259-equivalent residue was observed in vivo in mouse ACC²⁴, and Ser1259 is also partially phosphorylated in the insect cellexpressed human ACC used here. Size-exclusion chromatography coupled to multi-angle laser light scattering of BRCT in the presence of mono-phosphorylated (1255-CFSDSPPQ-pS-PTFPEAG-1270) and di-phosphorylated (1255-CFSD-pS-PPQ-pS-PTFPEAG-1270) ACC peptides provides supporting evidence for a dimeric BRCT-di-phosphopeptide complex (Extended Data Fig. 8h). This propensity for the formation of dimeric complexes enables BRCT to act as a molecular clamp for interlinking ACC dimers in ACC-BRCT and explains the dependence of formation of this filament type on BRCT. Interaction with full-length BRCA1 inhibits dephosphorylation of pSer80 in ACC¹³. This effect is not directly explained by the binding of dimeric BRCT, which contains only 213 of the 1863 amino acids of BRCA1. However, the minimal distance between BRCT and the pSer80-containing BC domains in ACC-BRCT is only 40 Å, and is compatible with sequestration of pSer80 by other regions of BRCA1 after BRCT recruitment.

The cryo-EM reconstructions reveal how ACC activity is regulated by formation of distinct filament types (Fig. 5), although the binding of the activator citrate itself cannot be visualized at the current resolution. Non-polymeric ACC dimers sample catalytically competent and incompetent states. Citrate locks ACC in a catalytically competent state by inducing formation of ACC-citrate filaments (Fig. 5). Notably, the non-polymerizing yeast ACC is also regulated by conformational locking: Phosphorylation at Ser1157 and subsequent binding of pSer1157 to a positively charged crevice in the CD disfavours formation of the active, closed dimer^{11,12,22}. Citrate binding in human ACC might also exert its effect via the CD, although binding at the BC domain dimer interface has also previously been suggested²⁵. The inhibitor palmitoyl-CoA modifies citrate-induced filaments, presumably resulting in reversible release of BC domain dimerization. AMPK-mediated Ser80 phosphorylation has also been proposed to disturb BC domain dimerization^{12,22}.

BRCA1 has been implicated in cell cycle-dependent regulation of fatty acid biosynthesis and lipogenesis in adipose tissue by interaction with ACC¹³. Our results extend prior work on the interaction of cytosolic ACC with BRCT domains of BRCA1, which is primarily localized to the nucleus, but also occurs in the cytosol²⁶. The interaction of dimerizing BRCT domains with ACC phosphosite loops induces polymerization of open, extended ACC dimers into double-stranded ACC–BRCT filaments with a distinct CD conformation and monomeric BC domains (Fig. 5). BRCA1 has been reported to maintain ACC in an inactivated state by preventing dephosphorylation of pSer80²⁷. In the ACC–BRCT filament, the corresponding interaction could not be visualized, but it may nevertheless be explained by the resultant proximity of full-length BRCA1 to an exposed Ser80 in the monomeric BC domain. Further studies will be required to confirm and define the in vivo interplay of ACC and BRCA1 and the consequences for the regulation of both proteins.

ACC is a textbook example of the formation of regulatory filaments of metabolic enzymes²⁸. ACC has a central role in primary metabolism; its upregulation is linked to obesity-related diseases^{1,29,30} and tumour growth^{31–33}. Fungal ACC is also a target of the antifungal polyketide soraphen A¹⁹. By identifying distinct interactions in ACC filaments our data also provide a structural basis for manipulating ACC polymerization in vitro and in vivo, for example, for therapeutic intervention against ACC overactivation.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0201-4.

Received: 2 July 2017; Accepted: 24 April 2018; Published online 13 June 2018.

- Tong, L. Structure and function of biotin-dependent carboxylases. Cell. Mol. Life Sci. 70, 863–891 (2012).
- Wakil, S. J., Titchener, E. B. & Gibson, D. M. Evidence for the participation of biotin in the enzymic synthesis of fatty acids. *Biochim. Biophys. Acta* 29, 225–226 (1958).
- Bianchi, A. et al. Identification of an isozymic form of acetyl-CoA carboxylase. J. Biol. Chem. 265, 1502–1509 (1990).



- Ha, J., Daniel, S., Broyles, S. S. & Kim, K. H. Critical phosphorylation sites for acetyl-CoA carboxylase activity. J. Biol. Chem. 269, 22162-22168 (1994).
- Brownsey, R. W., Boone, A. N., Elliott, J. E., Kulpa, J. E. & Lee, W. M. Regulation of 5. acetyl-CoA carboxylase. Biochem. Soc. Trans. 34, 223-227 (2006).
- Vagelos, P. R., Alberts, A. W. & Martin, D. B. Activation of acetyl-CoA carboxylase and associated alteration of sedimentation characteristics of the enzyme. Biochem. Biophys. Res. Commun. 8, 4-8 (1962).
- Kleinschmidt, A. K., Moss, J. & Lane, D. M. Acetyl coenzyme A carboxylase: filamentous nature of the animal enzymes. Science 166, 1276-1278 (1969).
- Moss, J. & Lane, M. D. Acetyl coenzyme A carboxylase. IV. Biotinyl prosthetic group-independent malonyl coenzyme A decarboxylation and carbosyl transfer: generalization to other biotin enzymes. J. Biol. Chem. **247**, 4952–4959 (1972)
- Meredith, M. J. & Lane, M. D. Acetyl-CoA carboxylase. Evidence for polymeric filament to protomer transition in the intact avian liver cell. J. Biol. Chem. 253, 3381–3383 (1978).
- 10. Ashcraft, B. A., Fillers, W. S., Augustine, S. L. & Clarke, S. D. Polymer-protomer transition of acetyl-CoA carboxylase occurs in vivo and varies with nutritional conditions. J. Biol. Chem. 255, 10033-10035 (1980).
- Hunkeler, M., Stuttfeld, E., Hagmann, A., Imseng, S. & Maier, T. The dynamic organization of fungal acetyl-CoA carboxylase. *Nat. Commun.* 7, 11196 (2016).
- 12. Wei, J. & Tong, L. Crystal structure of the 500-kDa yeast acetyl-CoA carboxylase holoenzyme dimer. Nature **526**, 723–727 (2015).
- Ray, H., Suau, F., Vincent, A. & Dalla Venezia, N. Cell cycle regulation of the BRCA1/acetyl-CoA-carboxylase complex. *Biochem. Biophys. Res. Commun.* **378**, 13. 615-619 (2009)
- 14. Magnard, C. et al. BRCA1 interacts with acetyl-CoA carboxylase through its tandem of BRCT domains. Oncogene 21, 6729–6739 (2002).
- Shen, Y. & Tong, L. Structural evidence for direct interactions between the BRCT domains of human BRCA1 and a phospho-peptide from human ACC1. Biochemistry **47**, 5767–5773 (2008).
- Kim, C. W. et al. Induced polymerization of mammalian acetyl-CoA carboxylase by MIG12 provides a tertiary level of regulation of fatty acid synthesis. Proc. Natl Acad. Sci. USA **107**, 9626–9631 (2010).
- 17. Shen, Q. J. et al. Filamentation of Metabolic Enzymes in Saccharomyces cerevisiae. J. Genet. Genomics 43, 393-404 (2016).
- Suresh, H. G. et al. Prolonged starvation drives reversible sequestration of lipid biosynthetic enzymes and organelle reorganization in Saccharomyces cerevisiae. Mol. Biol. Cell 26, 1601-1615 (2015)
- 19. Shen, Y., Volrath, S. L., Weatherly, S. C., Elich, T. D. & Tong, L. A mechanism for the potent inhibition of eukaryotic acetyl-coenzyme A carboxylase by soraphen A, a macrocyclic polyketide natural product. Mol. Cell 16, 881-891 (2004).
- 20. Harriman, G. et al. Acetyl-CoA carboxylase inhibition by ND-630 reduces hepatic steatosis, improves insulin sensitivity, and modulates dyslipidemia in rats. Proc. Natl Acad. Sci. USA 113, E1796-E1805 (2016).
- 21. Madauss, K. P. et al. The human ACC2 CT-domain C-terminus is required for full functionality and has a novel twist. Acta Crystallogr. D 65, 449–461 (2009)
- 22. Wei, J. et al. A unified molecular mechanism for the regulation of acetyl-CoA carboxylase by phosphorylation. Cell Discov. 2, 16044 (2016).
- 23. Wu, O. et al. Structure of BRCA1-BRCT/Abraxas complex reveals phosphorylation-dependent BRCT dimerization at DNA damage sites. Mol. Cell **61**, 434-448 (2016).
- Huttlin, E. L. et al. A tissue-specific atlas of mouse protein phosphorylation and expression. Cell 143, 1174–1189 (2010).
- Kwon, S. J., Cho, Y. S. & Heo, Y. S. Structural insights into the regulation of ACC2
- by citrate. *Bull. Korean Chem. Soc.* **34**, 565–568 (2013).

 26. Thompson, M. E. BRCA1 16 years later: nuclear import and export processes. FEBS J. **277**, 3072–3078 (2010).
- Moreau, K. et al. BRCA1 affects lipid synthesis through its interaction with acetyl-CoA carboxylase. J. Biol. Chem. 281, 3172–3181 (2006).

- 28. O'Connell, J. D., Zhao, A., Ellington, A. D. & Marcotte, E. M. Dynamic reorganization of metabolic enzymes into intracellular bodies. Annu. Rev. Cell Dev. Biol. 28, 89-111 (2012).
- Harwood, H. J. Jr. Acetyl-CoA carboxylase inhibition for the treatment of metabolic syndrome. Curr. Opin. Investig. Drugs 5, 283-289 (2004)
- Stiede, K. et al. Acetyl-coenzyme A carboxylase inhibition reduces de novo lipogenesis in overweight male subjects: a randomized, double-blind, crossover study. Hepatology 66, 324-334 (2017).
- Guri, Y. et al. mTORC2 promotes tumorigenesis via lipid synthesis. Cancer Cell **32**, 807–823 (2017).
- Swinnen, J. V., Brusselmans, K. & Verhoeven, G. Increased lipogenesis in cancer cells: new players, novel targets. Curr. Opin. Clin. Nutr. Metab. Care 9, 358-365
- Svensson, R. U. et al. Inhibition of acetyl-CoA carboxylase suppresses fatty acid synthesis and tumor growth of non-small-cell lung cancer in preclinical models. Nat. Med. 22, 1108-1119 (2016).

Acknowledgements We thank T. Sharpe at the Biophysics facility, A. Schmidt at the Proteomics Core Facility, and the Imaging Core Facility, in particular A. Ferrand, of the Biozentrum Basel for protein characterization and imaging support, and EMBL Heidelberg for providing the pETG-10A vector. We thank sciCORE at University of Basel for support with high performance computing. A.H. is supported by a Fellowship for Excellence from the Biozentrum Basel International PhD program. M.H. was supported by a Novartis Excellence Fellowship. This work was supported by Swiss National Science Foundation grants 138262, 159696 and 164074.

Reviewer information Nature thanks R. Haselkorn, J. Kollman, M. St. Maurice and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.H. conceived the study, purified proteins, identified filaments by negative stain electron microscopy, prepared and optimized cryo-EM grids and collected data for ACC-BRCT, performed activity assays and MALS, processed ACC-BRCT data and refined the model, processed ACC-citrate data, interpreted data, prepared figures and wrote the manuscript. A.H. prepared, screened and optimized cryo-EM grids and collected data for ACC-BRCT and ACC-citrate, processed ACC-citrate data and refined the model, processed negative stain electron microscopy data, performed the streptavidin shift assay, interpreted data, prepared figures and wrote the manuscript. E.S. cloned ACC and BRCT and established purification procedures, built the model of the BT-CD region. Y.G. designed and executed experiments, analysed data. M.C. and H.S. contributed to electron microscopy sample preparation and data collection. T.M. conceived the study, interpreted data and wrote the manuscript. All authors critically reviewed the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0201-4.

Supplementary information is available for this paper at https://doi. org/10.1038/s41586-018-0201-4.

Reprints and permissions information is available at http://www.nature.com/ reprints

Correspondence and requests for materials should be addressed to M.H. or

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Protein expression and purification. The coding sequence for full-length ACC1 (amino acids 1-2346, Genebank accession #Q13085) was cloned into a modified pACEBACI (Geneva Biotech) expression vector containing a Gateway (LifeTechnologies) cassette with an N-terminal His₁₀-Myc-FLAG tag according to the manufacturer's instructions. Bacmid and virus generation was carried out in Sf21 cells (Expression Systems) in Insect-Xpress medium (Lonza), following the MultiBac instructions. No mycoplasma contamination was detected using the MycoAlert Mycoplasma Detection Kit (Lonza), no identification of cell lines was carried out. The cells were harvested three days after infection by centrifugation and stored at -80 °C until further use. Cell pellets were dissolved in lysis buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 40 mM imidazole, 2.5 mM MgCl₂, 5% glycerol, 5 mM β -mercaptoethanol (β -ME)) with the addition of a spatula tip of DNaseI, lysed by sonication and the lysate was cleared by ultracentrifugation. Soluble protein was purified by immobilized metal affinity chromatography using Ni-charged resin (Genscript, Ni-IMAC). After elution from the Ni-IMAC, the buffer was exchanged to gel filtration buffer (20 mM bicine pH 8, 150 mM NaCl, 5% Glycerol, 5~mM TCEP) using a Sephadex G-25 (GE Healthcare) column. To ensure highest biotinylation levels, ACC was biotinylated in vitro overnight in a reaction containing 50 mM Tris-HCl, pH 8.0, 16 mM bicine pH 8, 5.5 mM MgCl₂, 180 mM NaCl, 3 mM ATP, 14% glycerol, 4 mM TCEP, 0.5 mM biotin and $3.7\,\mu\text{M}$ Bir A. The degree of biotinylation was assessed by gel shift upon streptavidin binding (Extended Data Fig. 5d); 0.5 μg ACC was incubated at ratios of 2:1, 1:1, 1:2 and 1:5 with streptavidin (Sigma) for 30 min at room temperature. Samples were analysed by SDS-PAGE using a 4-15% Mini-Protean TGXTM Precast Gel (Biorad). The buffer was exchanged again, and ACC was concentrated using a 100,000 molecular mass cut off centrifugal concentrator (Amicon) and polished by size-exclusion chromatography (Superose 6, GE Healthcare). Dephosphorylated ACC was obtained after overnight incubation with λ -protein phosphatase (New England Biolabs) before the last gel filtration step. The removal of the phosphoryl groups to 92% completion was confirmed by mass spectrometry. Purified ACC was concentrated to 4 mg/ml and 2.7 mg/ml for phosphorylated and dephosphorylated protein, respectively, in gel filtration buffer.

Coding sequences for BRCT domains of human BRCA1 (Genebank accession #BC115037), delivered in pCR-BluntII-TOPO vector, were sub-cloned into pETG-10A (EMBL) and expressed in arabinose-inducible One Shot BL21-AI *Escherichia coli* cells (ThermoFisher, C607003) overnight at 16 °C. Cells were harvested, dissolved in lysis buffer (50 mM Tris-HCl pH 8.5, 150 mM NaCl, 25 mM imidazole, 2 mM MgCl₂, 5% glycerol, 5 mM β -ME) and lysed by sonication. The protein was purified by Ni-IMAC and size-exclusion chromatography (Superdex200, GE Healthcare) and concentrated to 7.8 mg/ml in ACC gel filtration buffer. Proteins were stored at $-80\,^{\circ}$ C after flash-freezing in liquid nitrogen.

ACC polymerization and negative stain electron microscopy. The small protein modifier MIG12 has been reported to be involved in ACC filament formation 16, however in our in vitro experiments with purified components, the presence of MIG12 was not necessary for the formation of the ACC-citrate or other filament forms. To form ACC-citrate filaments, dephosphorylated ACC was dialysed overnight against 50 mM Hepes/KOH pH 7.5, 10 mM K₃citrate, 0.1 mM EDTA, 5 mM β -ME and diluted to 20–40 µg/ml in the same buffer without β -ME. The same treatment was applied to phosphorylated protein, which yielded aberrant filaments and rings (Extended Data Fig. 1). In order to obtain ACC-BRCT filaments, phosphorylated ACC was mixed with an eightfold molar excess of BRCT domains and incubated at room temperature for 1 h, before dilution using gel filtration buffer to 25 µg/ml. The same treatment was applied to dephosphorylated protein, which did not result in filament formation (Extended Data Fig. 1). For ACC-citrate^{palm} filaments, ACC-citrate was prepared and palmitoyl-CoA was added in tenfold molar excess. Phosphorylated and dephosphorylated ACC was diluted using gel filtration buffer to 25–40 μg/ml to obtain the ACC protomer samples. All negative stain grids were prepared by applying 5 μl of protein solution to 200-mesh carbon-coated copper grids (prepared in house). The sample was allowed to adsorb for 5 s. Subsequently, the grids were washed three times with the corresponding buffer and once with water, before two rounds of staining using 2% uranyl acetate for 5 s and 20 s, respectively. Negative stain micrographs were acquired on a Philips CM-100 TEM at a nominal magnification of 92,000×.

Negative stain electron microscopy data collection and processing. Grids of human ACC protomers were prepared as described above and imaged on an FEI Talos TEM at 200 kV equipped with an FEI Ceta 16M Pixel CMOS camera. Data was collected at a nominal magnification of 73,000×, resulting in a pixel size of 2.01 Å. Contrast transfer function (CTF) was estimated using gCTF³⁴ from 251 collected images. Particles (22,005) were picked semi-automatically using boxer implemented in EMAN2³⁵. Iterative 2D classification was carried out using Relion³⁶ to yield 2D class averages containing 9,920 particles. Grids of ACC-citrate filaments were prepared as described above and imaged on an FEI Talos TEM at 200 kV equipped with an FEI Ceta 16M Pixel CMOS camera. Data was collected

at a nominal magnification of $57,000\times$, resulting in a pixel size of 2.59 Å. CTF was estimated using gCTF from 702 collected images. Particles (14,392) were picked manually using Relion. Iterative 2D classification was carried out using cryoSPARC³⁷ to yield 2D class averages consisting of 2,379 particles.

Cryo-EM grid preparation and data collection. For ACC–citrate filaments, dephosphorylated ACC was dialysed over night against 50 mM Hepes/KOH pH 7.5, 10 mM K_3 citrate, 0.1 mM EDTA, 5mM β -ME. For initial grids, the protein was diluted to 300 µg/ml using the same buffer without β -ME, and Lacey grids (Carbon Cu 300 mesh, Ted Pella) were prepared using a FEI Vitrobot Mark IV (4 °C, 3 s blotting time) with 4 µl sample. Optimized grids were prepared as described above, however, the protein was diluted to 400 µg/ml; n-dodecyl- β -D-maltoside was added to 17 µM and the mixture was incubated for 1 h at room temperature. To obtain ACC–BRCT filaments, phosphorylated ACC (4 mg/ml) was mixed with an eightfold molar excess of BRCT and dialysed against gel filtration buffer without glycerol. Filaments were diluted to 0.75 mg/ml directly before applying to the grids. Lacey grids (Carbon Cu 300 mesh, Ted Pella) were prepared using a FEI Vitrobot Mark IV (4 °C, 3.5 s blotting time).

Samples were imaged using an FEI Titan Krios equipped with a Gatan image filter (Quantum-LS GIF, 20 eV zero loss filtering) and a post-GIF K2 summit direct electron detector (Gatan). Images were recorded at 300 kV with a nominal magnification of $130,000\times$ in super-resolution mode with a pixel size of 0.529 Å per super-resolution pixel on the specimen level, applying a defocus range of $-1~\mu m$ to $-2.5~\mu m$ for ACC–citrate and -1 to $-3.5~\mu m$ for ACC–BRCT in dose fractionation mode. For ACC–citrate filaments, 40 frames, at $\sim 1~e^-$ Å $^{-2}$ per frame (yielding a total dose of 40 e $^-$ Å $^{-2}$) were recorded; for ACC–BRCT filaments 80 frames, at $\sim 1~e^-$ Å $^{-2}$ per frame with a total dose of 80 e $^-$ Å $^{-2}$ were recorded.

Image processing. For ACC-citrate filaments, 13,671 movies were recorded using SerialEM³⁸. Recorded movies were pre-processed online with FOCUS³⁹. Movies were Fourier-cropped from 8k to 4k, resulting in an effective pixel size of 1.058 Å, aligned and corrected for beam-induced motion using Motioncor2⁴⁰. CTF was estimated using CTFFIND4.141. Poor quality micrographs were rejected by CTF resolution estimation (>10 Å), sample drift (>80 Å) and by defocus values (< -0.5 and > -3.5) in FOCUS. Particles (247,337) were boxed manually using Relion 2.0.3 from 13,671 micrographs⁴², and all further processing steps were conducted in Relion 2.1b1, unless mentioned otherwise (Extended Data Fig. 2). 2D classification was performed using cryoSPARC and the resulting 174,224 particles were used for an ab-initio reconstruction. This volume was then used as a starting model for 3D classification in Relion. The volume of the class with the highest population was used for placement of ACC domain crystal structures (BC, PDB ID: 2YL2; CD, 5I87¹¹; CT, 4ASI) and subsequent mask creation. A masked 3D classification was performed and the two classes with the highest population were selected for a focused 3D refinement with 147,822 particles yielding a map at 6.6 Å resolution. Movie refinement, particle polishing, and subsequent additional 3D classification yielded a population of 131,062 particles, which were used for a final focused 3D refinement. Post processing yielded a map at 5.4 Å resolution as judged by FSC using the 0.143 threshold criterion⁴³. Local resolution was determined by ResMap using half-map reconstructions⁴⁴. Applying global C2 symmetry during reconstruction did not improve the overall map quality or resolution considerably. For ACC-BRCT filaments, 3,233 movies were recorded, aligned and corrected for beam-induced motion using Motioncor2⁴⁰ (Extended Data Fig. 3). Poor quality micrographs were rejected by CTF resolution estimation (gCTF³⁴, >6 Å), sample drift (>80 Å) and by defocus values (< -0.5 and > -3.5). Particles were boxed semi-automatically using Relion 2.0.1, and all further processing steps were conducted in Relion 2.0.1, unless mentioned otherwise (Extended Data Fig. 3). Unsupervised 2D classification was performed using 67,903 particles from 2,924 micrographs, and particles that did not contribute to high-resolution class averages were excluded from further refinement steps. An initial model was generated using e2initialmodel.py implemented in EMAN235, filtered to 50 Å and used for 3D classification into four classes. Particles from the two classes with the highest population (44,997) were used for 3D refinement, yielding a map at 9.7 Å. The quality of the map was considerably increased during post processing and movie refinement to 7.7 Å. This map was used to unambiguously dock the individual crystal structures listed above, yielding a first model that was used to create two masks. The first mask contained the core particle consisting of CT, CD and BT domains. The second mask contained four full nodes. The particles were re-classified with imposed C2 symmetry and a focused refinement using the core mask yielded a map at 6.6 Å. After post processing using the two different masks and movie refinement, the resolution of the maps was increased to 4.6 Å and 5.9 Å, respectively, based on the FSC 0.143 threshold criterion⁴³. The local resolution was determined by ResMap using half-map reconstructions⁴⁴.

Model building and refinement. Using the highest resolution maps of both filaments, crystal structures of ACC and BRCA1 BRCT domains (PDB ID: 4Y18²³) were manually placed and then rigid-body fitted into the maps using Chimera⁴⁵ and Coot⁴⁶. Post-map averaging based on additional local two-fold symmetry

combined with adjusted B factor sharpening further increases map quality (Extended Data Fig. 4), and has been used in model building, but not for final map representation. Local FSC calculations were done using the localfsc Chimera plugin⁴⁷. Applying complete domain-wise local symmetry operators during refinement in Relion did not yield substantially better maps or resolution.

An atomic model of human ACC was created, first for the higher resolution ACC-BRCT filament, based on the available high-resolution crystal structures of human ACC domains and a homology model of human ACC-BCCP based on yeast ACC-BCCP generated by SWISSMODEL⁴⁸. Loop regions were excised from or added to the model based on EM maps. The phosphorylated Ser1263 and the surrounding loop directly bound to the BRCT domains were modelled according to the ACC phosphopeptide-BRCT co-crystal structure (PDB ID: 3COJ¹⁵). All models were protonated and refined using rigid body refinement, gradient-driven minimization and simulated annealing refinement implemented in phenix.real_space_refine⁴⁹, employing NCS- and reference model restraints, followed by a final round of ADP refinement. The final model of ACC-citrate contains residues 102-511; 524-543; 556-1188; 1230-1256; 1284-1333; 1352-1518 and 1525-2338. ACC in the final ACC-BRCT filament model (mask 2) contains residues 102-267; 278-511; 524-543; 554-617; 625-707; 714-748; 752-821; 832-839; 848-1188; 1230-1256; 1261-1270; 1284-1333; 1352-1430; 1436-1549; 1554-1560 and 1564-2335. The BRCT domains of BRCA1 contain residues 1648-1859. The final ACC-BRCT model (mask 1) contains residues 625-707; 714-748; 832-839; 848-1188; 1230-1256; 1258-1333; 1352-1430; 1436-1550; 1554-1560 and 1565-2337. Data collection parameters and refinement statistics are summarized in Extended Data Table 1.

Interface areas were computed using PDBePisa server⁵⁰; the values obtained are only approximate values and provided as such, due to the uncertainty of sidechain positioning at the given resolution. Twist and rise of the filaments was determined using lsqkab implemented in the CCP4 suite⁵¹, and are provided as approximate values due to the inherent bending and twisting along assembled filaments observed in the EM micrographs. The ACC–BRCT filament could alternatively be described as a right-handed single-stranded helix with a helical twist of ~120° and a rise of 95 Å. Protein alignment was calculated using Muscle align⁵². Figures of models and EM maps were generated using PyMOL (Schrödinger), Chimera and ChimeraX^{45,53}.

Size-exclusion chromatography coupled to multi-angle light scattering. SEC-MALS measurements were performed for samples of 2 mg/ml BRCT without phosphopeptide, with mono-phosphorylated peptide (ACC_p1, 1255-CFSDSPPQpS-PTFPEAG-1270) and di-phosphorylated peptide (ACC_p2, 1255-CFSDpS-PPQ-pS-PTFPEAG-1270), at room temperature (26 °C) in gel filtration buffer using a GE Healthcare Superdex 75 increase 5/150 GL column on an Agilent 1260 HPLC; (phospho)peptides were obtained commercially by chemical synthesis (Genscript). Elution was monitored using an Agilent multi-wavelength absorbance detector, a Wyatt Heleos II 8+ multiangle light scattering detector and a Wyatt Optilab rEX differential refractive index detector. The column was equilibrated overnight in running buffer to obtain stable baseline signals from the detectors before data collection. Inter-detector delay volumes, band broadening corrections, and light-scattering detector normalization were calibrated using an injection of 2 mg/ml BSA solution (ThermoPierce) and standard protocols in ASTRA 6. Weight-averaged molar mass (Mw), elution concentration, and mass distributions of the samples were calculated using the ASTRA 6 software (Wyatt

Activity assays. The catalytic activity of ACC was measured by following the incorporation of radioactive $^{14}\mathrm{C}$ into acid-stable non-volatile product 54 . The reaction mixture contained 0.125 µg recombinant ACC (4.7 nM) in 50 mM Hepes-KOH, pH 7.5, 3 mM ATP, 6 mM MgCl₂, 50 mM NaH $^{14}\mathrm{CO}_3$ (specific activity 7.4 MBq mmol $^{-1}$), 1 mM acetyl-CoA, 8 mM K3citrate, 150 nM malonyl-CoA reductase (MCR) 55 and 0.5 mM NADPH. The specific activity of MCR towards the ACC product malonyl-CoA, under expense of NADPH, was exploited to sequester malonyl-CoA (a potent inhibitor of ACC) from the reaction mixture. The total reaction volume was 100 µl. The reaction mixture was incubated for 10

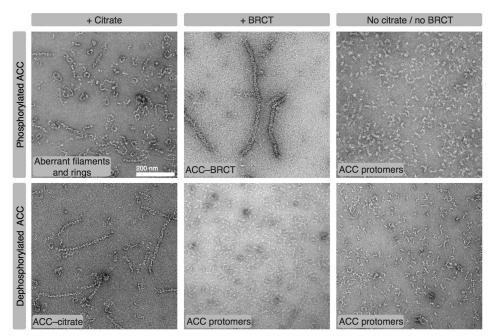
min at 32 °C, stopped by addition of 80 μl 6 M HCl and subsequently evaporated to dryness at 85 °C. The non-volatile residue was redissolved in 100 μl of water, 1 ml Ultima Gold XR scintillation medium (Perkin Elmer) was added, and the ^{14}C radioactivity was measured in a Packard Tricarb 2000CA liquid scintillation analyser. Measurements were carried out in three replicates and catalytic activities were calculated using a standard curve derived from measurements of varying concentrations of NaH $^{14}CO_3$ in reaction buffer.

Statistics and reproducibility. Polymerization of ACC into the distinct filaments was fully reproducible in all attempts. The streptavidin shift assay was performed four times and the SEC–MALS experiment was performed twice, after preceding size-exclusion chromatography analysis, with similar results.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

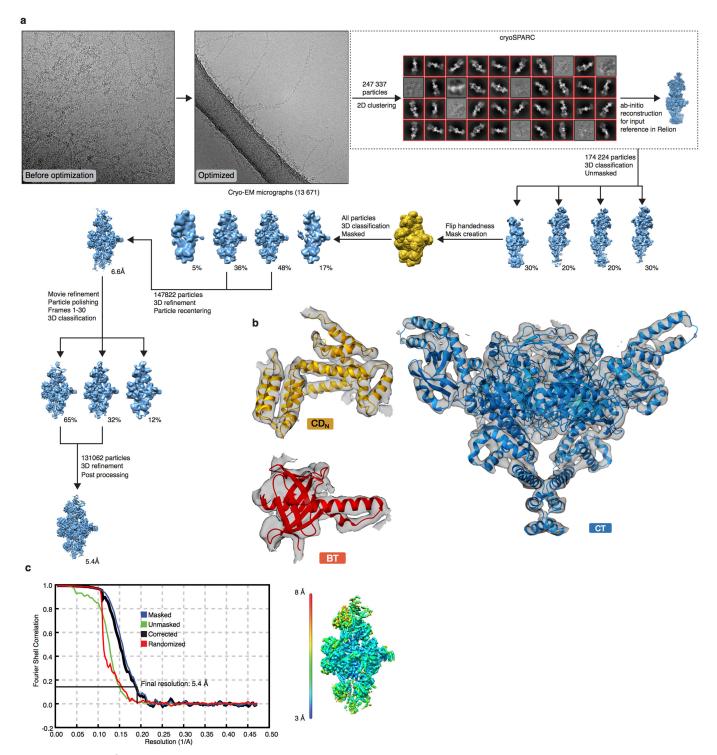
Data availability. The cryo-EM map of ACC-citrate has been deposited in the Electron Microscopy Data Bank as EMD-4342 and the corresponding model in the Protein Data Bank as PDB ID 6G2D. The cryo-EM maps of ACC-BRCT have been deposited in the EM Databank as EMD-4343 and EMD-4344 and the corresponding models in the Protein Data Bank as PDB ID 6G2H and 6G2I.

- Zhang, K. Gctf: Real-time CTF determination and correction. J. Struct. Biol. 193, 1–12 (2016).
- Tang, G. et al. EMAN2: An extensible image processing suite for electron microscopy. J. Struct. Biol. 157, 38–46 (2007).
- Scheres, S. H. W. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. J. Struct. Biol. 180, 519–530 (2012).
- Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* 14, 290–296 (2017).
- Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. J. Struct. Biol. 152, 36–51 (2005).
- Biyani, N. et al. Focus: The interface between data collection and data processing in cryo-EM. J. Struct. Biol. 198, 124–133 (2017).
- Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat. Methods 14, 331–332 (2017).
- 41. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
- Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. eLife 5, (2016).
- Scheres, S. H. W. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nat. Methods* 9, 853–854 (2012).
- Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* 11, 63–65 (2014).
- Pettersen, E. F. et al. UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612 (2004).
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. Acta Crystallogr. D 66, 486–501 (2010).
- Cardone, G., Heymann, J. B. & Steven, A. C. One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions. *J. Struct. Biol.* 184, 226–236 (2013).
- 48. Biasini, M. et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42**, W252–258 (2014)
- Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. D 66, 213–221 (2010).
- Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crytsalline state. J. Mol. Biol. 372, 774–797 (2007).
- Winn, M. D. et al. Overview of the CCP4 suite and current developments. Acta Crystallogr. D 67, 235–242 (2011).
- Crystallogr. D 67, 235–242 (2011).
 52. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797 (2004).
- Goddard, T. D. et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* 27, 14–25 (2018).
- Diacovich, L. et al. Kinetic and structural analysis of a new group of acyl-CoA carboxylases found in *Streptomyces coelicolor* A3(2). *J. Biol. Chem.* 277, 31228–31236 (2002).
- Kroeger, J. K., Zarzycki, J. & Fuchs, G. A spectrophotometric assay for measuring acetyl-coenzyme A carboxylase. *Anal. Biochem.* 411, 100–105 (2011).



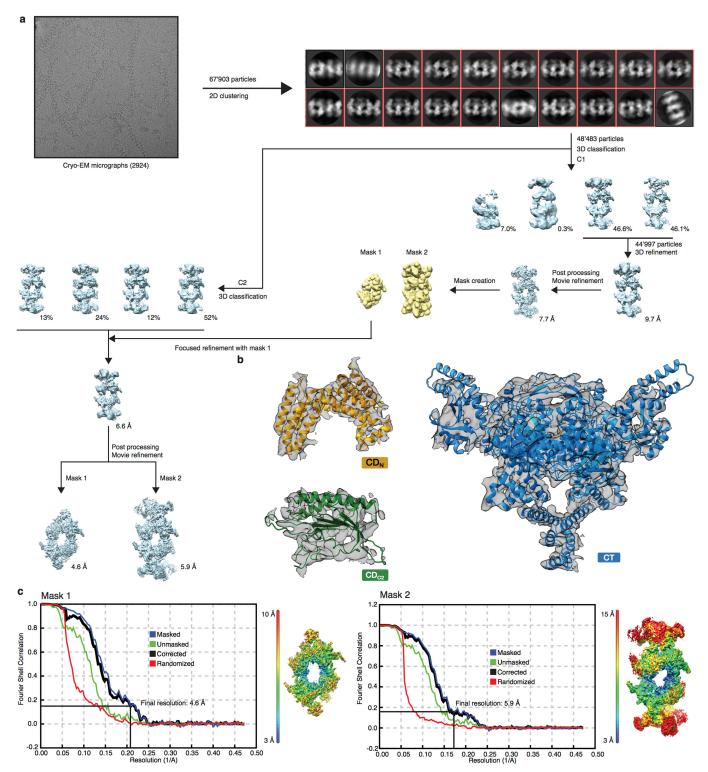
Extended Data Fig. 1 | Effect of citrate and BRCT domains on phosphorylated and dephosphorylated ACC. Negative stain electron microscopy micrographs of phosphorylated and dephosphorylated ACC in presence of citrate and BRCT. Addition of citrate to dephosphorylated ACC induces ACC-citrate filament formation, whereas when added to phosphorylated ACC, citrate results in aberrant ACC filament and ring

formation. Addition of BRCT domains to phosphorylated ACC induces formation of ACC–BRCT filaments, whereas no effect can be observed when adding BRCT domains to dephosphorylated ACC, and ACC remains in its dimeric form. In the absence of citrate or BRCT domains, phosphorylated and dephosphorylated ACC are in the dimeric, flexible form. Scale is identical across all images.



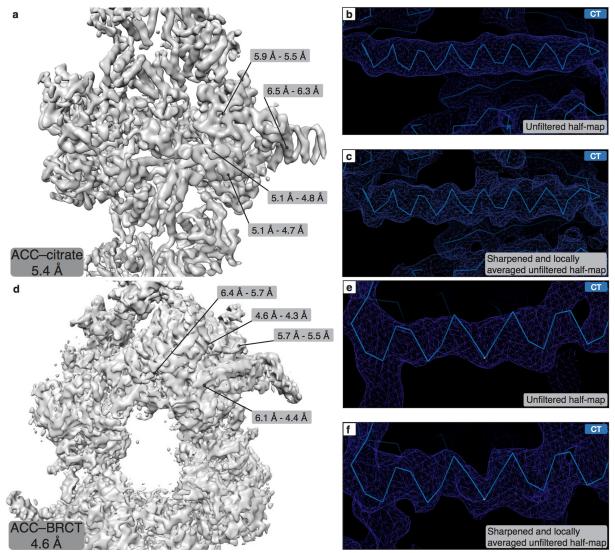
Extended Data Fig. 2 | Processing of electron microscopy data for ACC-citrate filaments and validation. a, Flowchart of data processing showing initial and optimized raw micrographs, 2D classes, 3D classes and refinement. Initial cryo-EM grids showed a meshwork of ACC-citrate filaments, exemplifying their flexible nature. After optimization, ACC-citrate filaments attach to carbon and protrude into holes. Some interaction between filaments can still be seen at the edge of the holes; however, single ACC-citrate filaments can clearly be recognized. 2D

classification and ab initio reconstruction were done in cryoSPARC, all other steps of processing were conducted in Relion. $\bf b$, Overview of map quality for the BT, CT and CD $_{\rm N}$ domains. Protein is shown in colour, according to the scheme in Fig. 1a, with transparent electron microscopy map. $\bf c$, FSC curves for masked and unmasked as well high-resolution phase-randomized reconstructions and final corrected FSC curve. Map coloured according to local resolution, colour scale is provided. All electron microscopy maps are shown at contour level of 0.0172.



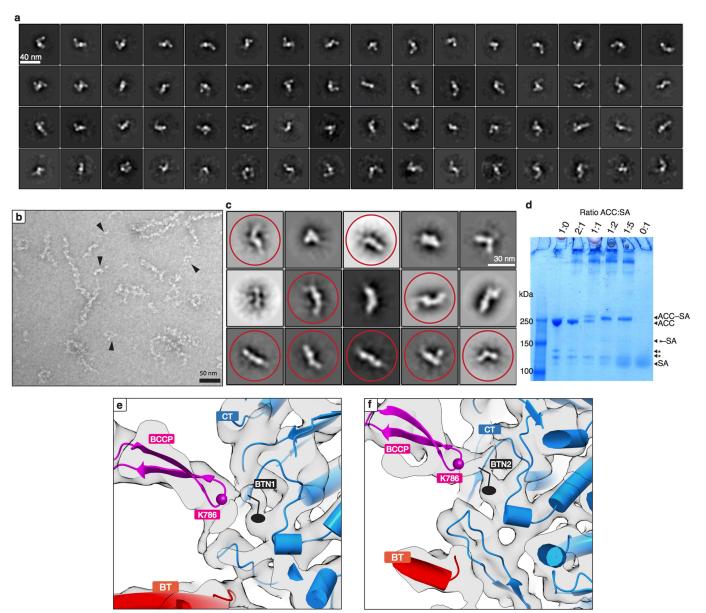
Extended Data Fig. 3 | Processing of electron microscopy data for ACC–BRCT filaments and validation. a, Flowchart of data processing showing a raw micrograph, 2D classes, 3D classes and refinement. b, Overview of map quality for the $CD_{C2},$ CT and CD_N domains. Protein is shown in colour according to scheme in Fig. 1a, with transparent electron microscopy map. Maps are shown at contour level of 0.009. c, FSC curves

for masked and unmasked as well high-resolution phase-randomized reconstructions and final corrected FSC curve. Maps coloured according to local resolution, colour scale is provided. Electron microscopy maps are shown at contour level of 0.019 and 0.105 for mask 1 and mask 2, respectively.



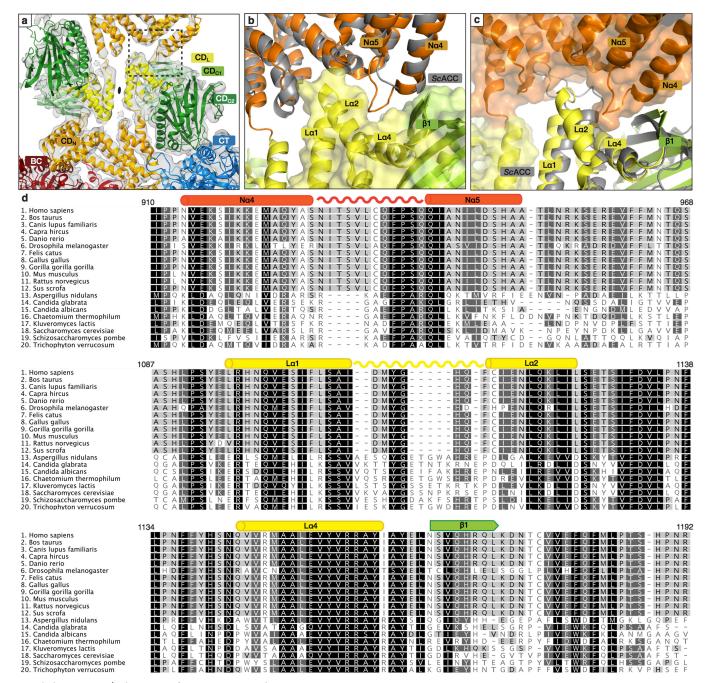
Extended Data Fig. 4 | Map improvement by local symmetry averaging. a, Overview map of ACC-citrate filament obtained after post processing in Relion (this map type is used for the other figures throughout the manuscript) at an overall resolution of 5.4 Å judged by FSC using the 0.143 threshold criterion. Labels indicate local resolution estimates for a box size of 40 pixel around the indicated positions, calculated using localfsc in Chimera using either the unfiltered half-maps directly (left value) or the unfiltered half-maps after local averaging as input (right value). Local resolution differentially improves after local averaging. Map is shown at contour level of 0.0172. b, Map around a helix in the CT domain in the ACC-citrate filament with density of unfiltered half-map as obtained directly from refinement in Relion. c, Image of the same helix in the CT domain with density of the same unfiltered half-map but after

local averaging for the CT domain. Here, additional B factor sharpening by $-120~\mbox{Å}^{-2}$ was applied before local averaging for visualizing additional detail (additional sharpening was not applied in ${\bf a}$ for localfsc comparison). Maps in ${\bf b}$ and ${\bf c}$ are shown at contour level of 0.012. ${\bf d}$, Overview map of ACC–BRCT filament, resolution values as obtained in ${\bf a}$ are indicated. Map is shown at contour level of 0.019. ${\bf e}$, Map around a helix in the CT domain in the ACC–BRCT filament with density of unfiltered half-map as obtained directly from refinement in Relion. ${\bf f}$, Image of the same helix in the CT domain with density of the same unfiltered half-map but after local averaging for the CT domain. Here, additional B factor sharpening by $-40~\mbox{Å}^{-2}$ was applied before local averaging for visualizing additional detail (additional sharpening was not applied in ${\bf d}$ for localfsc comparison). Maps in ${\bf d}$ and ${\bf f}$ are shown at contour level of 0.013.



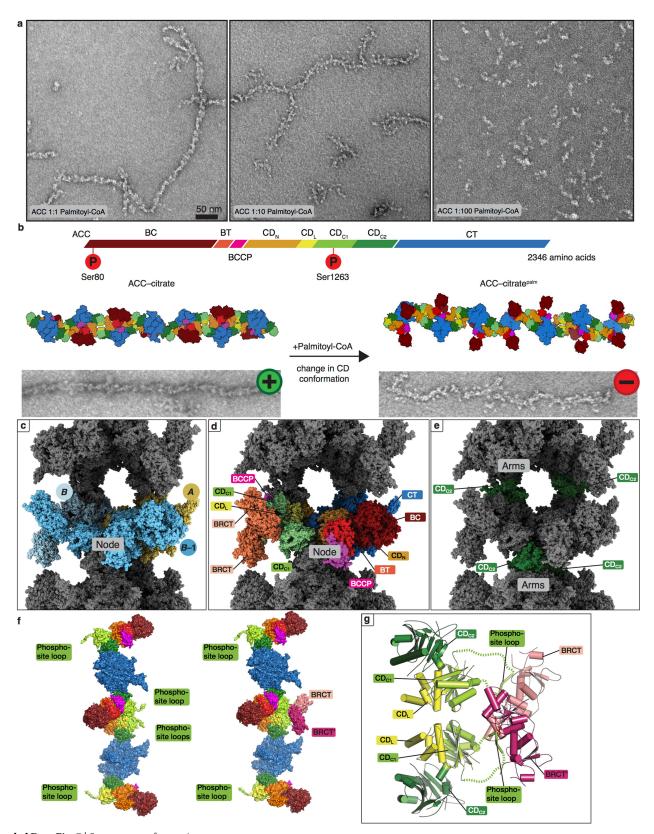
Extended Data Fig. 5 | Conformational variability of ACC dimers and density of the covalently linked biotin cofactor. a, Negative stain electron microscopy 2D class averages of dephosphorylated human ACC in the absence of citrate show the protein in a variety of conformations without considerable populations of closed dimers. b, Negative stain image of ACC–citrate filaments obtained from dephosphorylated ACC in a buffer with 10 mM citrate. Arrows indicate rarely observed, residual, non-polymerized ACC (further analysed in c). c, Negative stain electron microscopy 2D class averages of residual, non-polymerized ACC dimers observed in the presence of citrate on micrographs of polymerized ACC-citrate filaments (see b). ACC-like classes are marked in red. A variety of elongated conformations can be observed. d, Streptavidin (SA) shift assay to determine biotinylation level of ACC. ACC and streptavidin were mixed in different ratios and the shift upon

SDS-resistant binding of streptavidin to biotin was observed via SDS-PAGE. At higher excess of streptavidin, no unbound ACC is observed, indicating complete biotinylation of ACC. Two degradation products of ACC (indicated by asterisks) are observed, one of which also shows a band shift (*-SA) and thus contains the biotinylated site. Owing to the tetrameric nature of streptavidin, higher-order complexes are formed, which can also be observed on the gel. An uncropped image of the gel is shown in Supplementary Fig. 1. e, f, Density of the covalently linked biotin cofactor in the two active sites of the CT domain dimer. The main chain $C\alpha$ position of the biotinylated lysine (residue 786) is indicated. Due to limited resolution, the cofactor was not modelled; its orientation is shown schematically. For clarity, parts of BCCP are not shown. The map is shown at contour level of 0.0238.



Extended Data Fig. 6 | Alignment of CD sequences and ACC-citrate filament interface. a, The intermolecular interface in ACC-citrate filaments is shown in cartoon representation with the transparent electron microscopy map shown in grey. Local two-fold symmetry is indicated, and domains of the lower dimer are labelled. The map is shown at contour level of 0.0189. b, Close-up of the interface as indicated in a. ACC-citrate is shown in colour as cartoon representation. ScACC is shown as a cartoon in grey and superimposed for one side of the interface. For the other side, an additional surface representation is shown for ACC-citrate. The loop between N α 4 and N α 5 of the four-helix bundle of ACC-citrate CDN domain binds in the cradle formed by L α 2, L α 4 and β 1. This loop is substantially shorter in ScACC, demonstrating incompatibility of ScACC with formation of the interface. c, Same depiction as in b, but the surface of the top ACC-citrate dimer is shown and ScACC, shown in grey, is superimposed on the bottom dimer. The extended loop in ScACC between

Lα1 and Lα2 is not compatible with filament formation. **d**, Alignment of 20 ACC CD sequences of metazoan and fungal organisms. Residue numbers according to human ACC are indicated as well as the helices, the loops and the strand labelled in **b** and **c**. Darker colour indicates increased conservation. Pairwise identity over all aligned sequences is 61.5%, pairwise identity over metazoan and fungal sequences is 97.0% and 54.0%, respectively. Accession numbers: *Homo sapiens*, Q13085; *Bos taurus*, Q9TTS3; *Canis lupus familiaris*, E2RL01; *Capra hircus*, XP_017919660; *Danio rerio*, F1QH12; *Drosophila melanogaster*, Q7JV23; *Felis catus*, XP_011287256; *Gallus gallus*, P11029; *Gorilla gorilla gorilla*, XP_018881836; *Mus musculus*, Q5SWU9; *Rattus norvegicus*, P11497; *Sus scrofa*, D2D0D8; *Aspergillus nidulans*, AN6126.2; *Candida glabrata*, Q6FKK8; *Candida albicans*, C4YNG3; *Chaetomium thermophilum*, G0S3L5; *Kluveromyces lactis*, Q6CL34; *Saccharomyces cerevisiae*, Q00955; *Schizosaccharomyces pombe*, P78820; *Trichophyton verrucosum*, D4DIV5.

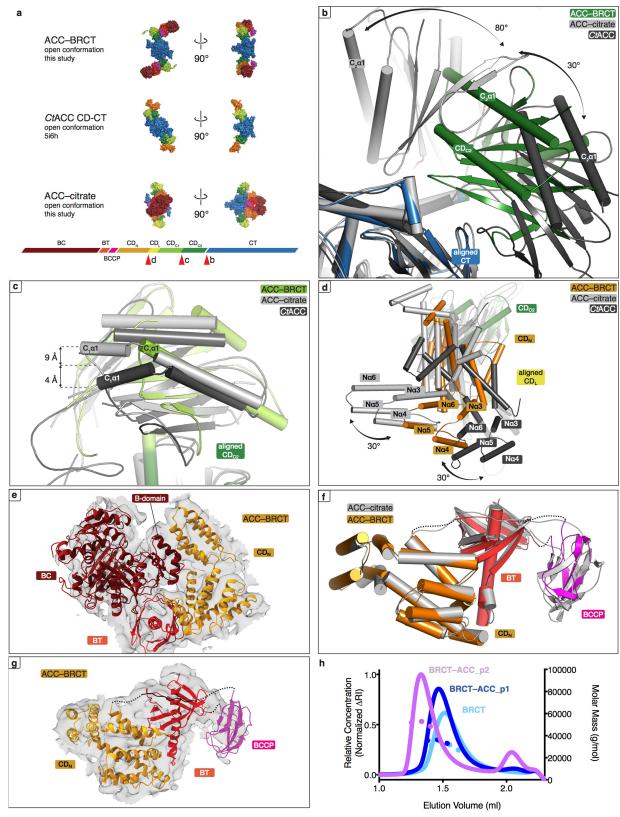


Extended Data Fig. 7 | See next page for caption.



Extended Data Fig. 7 | Impact of palmitoyl-CoA addition on ACC-citrate filaments and architecture of ACC-BRCT. a, Negative stain electron microscopy micrographs of ACC-citrate filaments treated with increasing concentrations of palmitoyl-CoA. At a 1:1 molar ratio of ACC-citrate monomer to palmitoyl-CoA, filaments show no differences to ACC-citrate filaments. At 1:10 molar ratio, ACC-citrate^{palm} filaments are observed. At 1:100 molar ratio, filaments dissolve. b, Top, domain organization of human ACC. Bottom left, enlarged negative stain electron micrograph of ACC-citrate filament with surface representation of the model coloured according to domains. Bottom right, electron micrograph of a ACC-citrate^{palm} filament and interpretation by a plausible model derived from ACC-citrate filaments by disrupting the BC domain dimers and flipping out of the BC domain. c, Surface representation of

ACC–BRCT with components of a single node coloured as in Fig. 3a. Domains of three molecules (A, B and B-1) add parts to the node. **d**, Same view as in **c**, but the domains are coloured according to the domain colour scheme in **b**. **e**, Same view as in **c**, with the CD_{C2} domains coloured according to domain colour scheme. These domains constitute the connecting arms between adjacent nodes. **f**, Surface representation of two consecutive dimers within one helix strand. Left, view without BRCT domains; the phosphosite loops are labelled. Right, view with dimeric BRCT domains establishing the connections between two dimers. **g**, Enlarged view of the phosphosite loop-BRCT interaction area, illustrating minimal contacts between the two CD_{C1} domains and between the filament strands and the BRCT domains. The interaction is governed by binding of the phosphosite loop to the dimeric BRCT.



Extended Data Fig. 8 | See next page for caption.



Extended Data Fig. 8 | Analysis of the architecture of ACC-BRCT filaments. a, Surface representation of dimers of ACC-BRCT, C. thermophilum (Ct) ACC CD-CT, and ACC-citrate, in the same relative orientation and coloured according to the sequence scheme shown below. **b**, CT-based overlay of the three structures, illustrating the rotations (indicated by arrows) of the CD_{C2} domains of ACC-citrate and CtACC relative to CD_{C2} of ACC-BRCT. CD-CT of ACC-BRCT, ACC-citrate and CtACC are shown in full colour, light grey and dark grey, respectively. Helix $C_2\alpha 1$ is labelled. c, CD_{C2} -based overlay of the three structures, representing the displacement (indicated) of CD_{C1} of ACC-citrate and CtACC relative to CD_{C1} of ACC-BRCT. Colouring as in **b**. CD_L was omitted for clarity, and helix $C_1\alpha 1$ is labelled. **d**, CD_L -based overlay of the three structures, illustrating the displacement of the CD_N of ACC-citrate and $\mathit{Ct}ACC$ relative to CD_N of ACC–BRCT. Colouring as in \boldsymbol{b} . The fourhelix bundle of helices $N\alpha 3$ – $N\alpha 6$ is labelled. The range of displacements of the ends of the bundle is indicated by arrows. e, BC, BT and CD_N domains

of ACC-BRCT filament together with the electron microscopy map at a low contour level to visualize less well-ordered regions and to illustrate the placement of the B-domain cap of the BC domain in the clamp-like CD_N domain. f, Overlay of CDN, BT and BCCP domains from ACC-BRCT and ACC-citrate, revealing a conserved conformation. g, CD_N, BT and BCCP domains in the ACC-BRCT filament are shown together with the electron microscopy map at low contour level to visualize the poorly ordered BCCP domain. Maps in e and g are shown at contour level of 0.007. h, The SEC-MALS elution profiles show the molecular mass (right axis) and the scattering intensity (Rayleigh ratio) at the 90° detector (left axis) of BRCT domains bound to the indicated ACC peptides. Elution for BRCT and the BRCT-ACC_p1 complex correspond to a mostly monomeric species with a dimeric subpopulation in fast equilibrium with an average molecular mass of 31.8 kDa and 34.8 kDa, respectively. The elution of BRCT-ACC_ p2 with a molecular mass of 51.1 kDa indicates a strong increase in the population of dimeric BRCT-peptide species.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	A C(C, C')	ACC DDCT 1	ACC DROTTA
	ACC-Cit	ACC-BRCT 1	ACC-BRCT 2
	(EMDB-4342)	(EMDB-4343)	(EMDB-4344)
	(PDB 6g2d)	(PDB 6g2h)	(PDB 6g2i)
Data collection and processing			
Magnification	130kx	130kx	130kx
Voltage (kV)	300	300	300
Electron exposure (e-/Ų)	40	80	80
Defocus range (μm)	-1 – -2.5	-13.5	-1 – -3.5
Pixel size (Å)*	1.058	1.058	1.058
Symmetry imposed	C1	C2	C2
Initial particle images (no.)	174224	67903	67903
Final particle images (no.)	131062	48483	48483
Map resolution (Å)	5.4	4.6	5.9
FSC threshold	0.143	0.143	0.143
Map resolution range (Å)	3 – 8	3 - 10	3 - 15
Refinement			
Initial model used (PDB code) [†]	Ab initio	Ab initio	Ab initio
Model resolution (Å)	<5.4 [‡]	4.7	6.1
FSC threshold	0.5	0.5	0.5
Model resolution range (Å)	5.4 - 476.1	4.6 - 397.8	5.9 – 397.8
Map sharpening B factor (\mathring{A}^2)	-266	-180	-202
Model composition	-200	-100	-202
Non-hydrogen atoms	44178	52044	145444
Protein residues	5501	6526	18352
Ligands	5501	-	\$
B factors (\mathring{A}^2)			
Protein	204.2	155.1	325.6
Ligand	204.2	155.1	\$25.0
R.m.s. deviations	_	_	
Bond lengths (Å)	0.0	0.0	0.0
Bond angles (°)	2.0	2.0	2.2
Validation	2.0	2.0	2.2
MolProbity score	2.35	2.40	2.20
Clashscore	3.5	4.5	5.6
Poor rotamers (%)	3.1	2.0	1.0
Ramachandran plot	J.1	2.0	1.0
Favored (%)	89.1	90.9	92.5
Allowed (%)	10.2	8.3	6.9
Disallowed (%)	0.6	0.7	0.6
Disallowed (70)	0.0	U./	0.0

^{*}Final pixel size after Fourier cropping.

†The following models were used for initiating model building: ACC-citrate: 2YL2; 5l87; 4ASI; ACC-BRCT 1: 5l87; 4ASI; ACC-BRCT 2: 2YL2; 5l87; 4ASI; 4Y18.

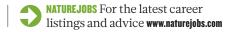
[‡]Masked cc in shell 5.41–5.35 is 0.61.

 $^{{}^{\}S}pSer1263$ is treated as a residue in refinement and not as ligand, therefore B factors are not calculated independently.

CAREERS

DEFORESTATION DETECTIVE Embracing industry to quell emissions **p.477**

BLOG Personal stories and careers counsel http://blogs.nature.com/naturejobs





Hobbies can be more than just a welcome distraction from the complexities of a scientific career — they bring many other benefits.

WORK-LIFE BALANCE

Play time for researchers

How hobbies can boost scientists' productivity and creativity.

BY JULIA ROSEN

Then Audrey Kelly isn't catching toads and analysing their DNA to study how species hybridize, she makes bread. Kelly is a fifth-year PhD student at the University of North Carolina, Chapel Hill, and she learnt to bake from her father before she moved away for her undergraduate programme. "It's kind of like a science experiment," she says, "but you get to eat it at the end."

As she does in the laboratory, Kelly records her methods and results in a notebook, but she doesn't take it too seriously: the hobby offers a break from the stress of doing science. "You're not as worried about screwing up," she says. "Your career's not on the line."

Many scientists struggle to take time away from the never-ending demands of research — and to flout the pervasive culture of overwork — to pursue personal interests.

Kelly says that she sometimes feels guilty when elbow-deep in dough. "If you are not working yourself to the bone and crying yourself to sleep every night," she asks rhetorically, "are you working hard enough?" Many surveys reveal that academic scientists regularly put in overtime; one poll, conducted by *Nature* in 2016, found that more than one-third of early-career researchers worked for more than 60 hours a week (see *Nature* 538, 446–449; 2016).

But evidence suggests that hobbies such as Kelly's nourish more than just the belly. Research has linked participation in leisure activities to many measures of physical and mental well-being, from reduced blood pressure to a sense of belonging^{1,2}. Many scientists say that their hobbies provide them with crucial opportunities to relax, to find satisfaction in completing small, defined projects and, occasionally, to make the kinds

of insightful leaps that propel science forward.

The key is to stop feeling bad about having interests outside research, says Alex Clark, associate vice-president of research at the University of Alberta in Edmonton, Canada, and co-author of a book called *How to be a Happy Academic* (2018). "We need to stop seeing hobbies and work as zero-sum games," he says.

PICKING A PASTIME

The first step in cultivating a hobby is to deliberately set aside personal time in which to pursue it, says Bailey Sousa, Clark's co-author and director of the International Institute for Qualitative Methodology at the University of Alberta. "We have 1,440 minutes in a day, and if we don't have control of that time," she says, "someone else is going to."

Carving out free time can be challenging, but there are strategies for success. Jingmei Li, a cancer researcher at the Genome Institute

CAREERS

of Singapore, uses most of her holiday time for diving trips. She had always feared swimming in the open ocean, but decided to face down her fear and got her scuba certificate ten years ago. Now, she's hooked and goes diving around the world.

Edward Davis, a palaeobiologist at the University of Oregon in Eugene, schedules time each week in which he tries not to work. He prefers to leave this time unstructured and follow his interests. His aim is "to not feel as though I have a bunch of additional goals I have to set for myself". He spends up to ten hours per week on hobbies, and explained in a 2013 blogpost that he found a good work-life balance when maintaining three extracurricular interests.

Jennifer Hertzberg, a palaeoceanographer at the Old Dominion University in Norfolk, Virginia, tries to do errands such as laundry and grocery shopping during the week. That way, she can keep the weekends free for her hobbies, which include doing jigsaw puzzles and assembling miniature Lego kits of birds and other animals. "I've amassed this whole collection; it's like a zoo," she says.

Hertzberg has enjoyed building with Lego since she was a child, and Clark says that looking back at childhood interests is a great place to start when seeking a hobby. "It may have been something you did when you were younger, but that fell to the wayside as you got sucked into a scientific career," Clark says. Alternatively, he recommends an activity to share with partners, friends and children.

Dean Simonton, a retired psychologist at the University of California, Davis, suggests that researchers analyse their own dispositions for clues about which pastimes might be the best fit for them. Those who are verbal thinkers, he says, might most enjoy reading, while those who think in visual terms might prefer something like painting or photography.

Signing up for a short course, class or workshop, or joining a group such as a choir, offers ways to explore new hobbies. A set schedule of meetings or rehearsals demands that scientists make time for the activities. Davis's newest interest is making custom knives for himself and to give as gifts. He belongs to a local group for enthusiasts, and attends meetings and workshops to hone his skills. The club also helps him to expand his social network. "I have more and different people that I'm talking to, and more and different ideas I'm exposed to," he says.

REAPING REWARDS

One of the biggest benefits for scientists in pursuing hobbies is that they give the mind a rest from the rigours of research. Everyone — including scientists — can get stuck in ways of thinking that prevent them from finding a solution to a vexing problem, no matter how hard they work. "You can't get outside the box you're in," says Simonton. Evidence suggests that taking a break³ or doing something different⁴ can help to weaken those associations and improve problem-solving, by revealing a



Theoretical physicist Nadav Drukker does pottery in his spare time.

new approach or an overlooked detail⁴.

One option for mentally unplugging is exercise. Davis says that almost every good idea he's had came to him while he was working out and letting his mind wander. And that's not the only benefit. Davis dabbled in sport as a youth, but got serious about swimming during his PhD programme at the University of California, Berkeley. "I noticed that my health was flagging as a consequence of the stress that I was under," he says. Exercising improved his fitness levels and helped to both mitigate his chronic asthma and relieve work-related pressure. "Vigorous

exercise helps burn some of the stress hormones that are produced when you are worrying about being successful," he says. It might also

"We need to stop seeing hobbies and work as zero-sum games."

have helped him to become a better researcher; studies suggest exercise boosts various brain functions⁵.

Eventually, Davis decided that he wanted to do triathlons, so he began to read books on becoming a runner. He learnt how to progress safely and efficiently as a novice, but also noticed that he had begun to feel more competitive. When this happens with a hobby, he says, it can add pressure instead of relieving it. So he made a conscious decision to pursue his own internal goals and not to compare himself against other competitors.

Nadav Drukker, who studies string theory at King's College London, relishes the opportunity to improve at pottery. Having mastered the basics, Drukker now makes pieces inspired by his work and decorated with equations. Last year, he had a solo exhibition at a London gallery that nearly sold out. He says that the pieces combine his love of physics with his love of ceramics, and provide a unique way to share his highly theoretical research with broader audiences.

Doing pottery also offers a break from research, which was what initially drew Drukker to the activity in graduate school. His work mostly involves "sitting in front of a calculation that I don't know how to solve", he says.

Eventually he realized that taking time off from physics increased his productivity. He says that shaping clay on the wheel gives his brain a rest when he starts going in circles, and that he finds the activity almost meditative.

Hobbies can also provide a sense of accomplishment when researchers are stuck in a rut at work. Hertzberg thinks that handling the tiny Lego blocks has helped her in the lab, where she uses tweezers to pick microscopic fossils out of ocean sediments for analysis. ("I have a steady hand," she says.) But what she likes most about her puzzles and Lego kits is that she can complete them in a relatively short amount of time while she labours on longterm scientific projects. "It's sort of like instant gratification," she says.

INSPIRATIONAL INTERESTS

Beyond offering the brain a breather, pastimes can also lead to inspiration. Simonton, who studies genius and creativity, says that reading about Buddhist philosophy influenced Nobel-prizewinning physicist Murray Gell-Mann's theory of subatomic particles. And when astronomers first caught a glimpse of the Moon through a telescope, Galileo quickly realized that the shadows on its face indicated that its surface was rough and mountainous — not smooth, as Aristotle had believed. That's because Galileo had dabbled in painting, and had learnt how to represent 3D objects on a flat canvas.

Often, the synergy of ideas happens by accident, Simonton notes. "It's not obvious," he says, "that an interest in painting would be useful to an astronomer." But to set the stage for serendipity to strike, researchers must assemble a broad set of knowledge and experiences to pull from. Any hobby helps, but the easiest option is to read material outside the scientific papers in one's discipline, says Simonton, who has an upcoming book on creativity called The Genius Checklist (2018).

Asking the brain to do different activities also builds cognitive flexibility, Simonton adds. Li says that diving has taught her greater attention to detail. Spotting well-camouflaged creatures under water requires heightened senses

and the ability to see the unexpected — just like scouring data for new insights. "The answer is right there," she says. "One just needs to have the right frame of mind to see it."

Despite the clear benefits of hobbies, however, they aren't always valued in the culture of science. "People actually hide their hobbies, or pretend they don't do anything outside of work, because they are worried about what people will think," says Sousa. But that's starting to change. For instance, the UK Academy of Medical Sciences launched its MedSciLife campaign in 2017 to highlight researchers who cook, craft and engage in all kinds of other non-academic activities. Social media has made it easier than ever for researchers to share personal interests.

Clark says that senior scientists can serve as role models and help to boost the acceptability of pastimes by making their own hobbies part of their professional identity. "That broadcasts important cultural signals that success in science and having a life need not be incompatible," he says. In fact, Clark argues, whereas researchers feel pressure to publish often, their legacy depends more on the quality — not the quantity — of their work. "That compels us to think about what makes us best placed to make the best contributions," he says. "And really, that is a way of living that is focused on creativity, innovation, vibrancy — and not on just producing more." ■

Julia Rosen *is a freelance writer in Portland, Oregon.*

- 1. Newman, D. B., Tay, L. & Diener, E. *J. Happiness Stud.* **15**, 555–578 (2014).
- Eschleman, K. J., Madsen, J., Alarcon, G. & Barelka, A. J. Occup. Organ. Psychol. 87, 579–598 (2014).
- 3. Sio, U. N. & Ormerod, T. C. Psychol. Bull. **135**, 94–120 (2009).
- 4. Baird, B. et al. Psychol. Sci. 23, 1117–1122 (2012).
- Prakash, R. S., Voss, M. W., Erickson, K. I. & Kramer, A. F. Annu. Rev. 66, 769–797 (2014).

CORRECTIONS

The Careers Feature 'Crunch time for data' (*Nature* **557**, 745–747; 2018) erroneously stated that an image from Planet was unavailable owing to a security concern. In fact, the reason for its unavailability was not specified. Also, DigitalGlobe is headquartered in Westminster, Colorado, not in Boulder.

The Careers Feature 'It takes more than a vow' (*Nature* **558**, 149–151; 2018) erroneously stated that Dorceta Taylor is director of diversity, equity and inclusion for the whole of the University of Michigan. In fact, she is head of these affairs just for the university's School for Environment and Sustainability.

BACK STORYDeforestation detective

Ecologist Lahiru Wijedasa at the National University of Singapore submitted a paper in 2015 that warned of future dangerous carbon emissions from Indonesia's peatland forests. The paper was finally published this month (L. S. Wijedasa et al. Glob. Change Biol. http://doi.org/cqtm; 2018). Wijedasa explains how his views changed during the process.

Why do peatland forests matter globally?

Peatland forests are carbon-rich swamps that have formed over centuries. In Indonesia, massive areas have been drained to grow crops, particularly oil palm and acacia. In 2011, the Indonesian government imposed a moratorium on issuing licences to clear land for industrial-scale development. But in 2015, fires on cleared lands produced more emissions than did the whole of Europe. Indonesia now has a Peatland Restoration Agency, which reports to the president and is mandated to restore 2 million hectares of peat forest by 2020. Our paper shows, however, that 51% of emissions will come from areas that have already been drained and are used for industrial agriculture.

That's bleak. What is the take-home message?

First, we need to maintain our remaining intact forest, of which 45% is not in protected or moratorium areas. My data show that 48% of the moratorium area isn't even peat swamp forest. Second, we'll need alternative forms of agriculture, so that communities can grow crops on wet peat soils.

Why did it take 3 years to publish your paper?

I submitted the paper in 2015. We went through four rounds of review and redid a lot; for example, we initially had three emissions scenarios, but increased those to the 18 defined by the Intergovernmental Panel on Climate Change. However, it was eventually rejected on the grounds of insufficient novelty. We then submitted it to *Global Change Biology*, which published it within three months.

Were your predictions higher than expected?

Data on peatland emissions have been controversial — in part, because some industry-funded studies have generated lower numbers. To address all potential scenarios, we assessed land-cover change from 1990 to 2010 using LandSat satellite imagery. Then we estimated emissions from peat between 1990 and 2130 for a range of agricultural expansions.



How did your views change?

Initially, I had thought that big palm-oil and acacia companies were solely to blame. But after spending more time in Sumatra and other areas of Indonesia, I saw that many of the company-owned forests are among the better-managed areas. Also, some of the palm-oil and acacia companies have set aside prime land for conservation, and have lobbied the government to protect forest that they legally could have developed. I now think that companies are part of the solution.

Did you consider community farmers?

Yes. Smallholders accounted for 60% of conversion outside the original government-designated areas. Whereas I might once have argued to restore all peatlands, I now better understand how much smallholders depend on the land, and that they clear forest to improve their livelihoods. Finding opportunities for sustainable agriculture could eliminate 51% of future emissions.

Does your work let palm-oil and acacia companies off the hook?

No. There are good companies and terrible companies, but the few companies who step up to work with the government are often the targets of bad press. Good companies are the best potential partners in conservation because they have the finances, enforcement ability and motivation — owing to public opinion — to protect these lands. And company-driven conservation has worked several times in Indonesia. It also offers a way for firms to atone for past deforestation in a country that desperately needs that help. ■

INTERVIEW BY VIRGINIA GEWIN

This interview has been edited for clarity and length.

GOING BACK FOR HITLER

The perils of time travel.

BY GEORGE NIKOLOPOULOS

Grail of time travellers. Every now and again some determined soul goes back in time to try, but all fail. It seems that Lady Time is fiercely protective of her strands and our actions are but proverbial ripples in the pond.

I may be young and frowned upon by my supposedly more knowledgeable colleagues, but I sincerely believe I have found the solution in Doctor Jablonski's papers. Entropy diligently thwarts those who actively attempt to change the past, he proposed, but if one goes with the flow, so to speak, they might be able to act in small ways, thereby precipitating big changes.

As most time travellers realize, this is clearly an impossible situation. If you go back in time with the express purpose of changing the past, you will most definitely do it willingly. That's where my own *humble* contribution lies, and that's why I will succeed where everyone else has failed. I'm just about to erase my memories and travel back in time innocent as a newborn babe.

I have read everything about the monster. I know all his crimes and all his atrocities by heart; you could even say that I have brainwashed myself. It's all burnt into my mind, so deep that even after wiping my memory clean some vestige will remain, buried in the recesses of my unconscious. I won't even know that I come from the 'future', but I will be subconsciously drawn to him and — although oblivious of the reason — I will hate him enough to kill him.

I know the price to pay; unaware that there's a *present* to go back to, I will remain stranded in the past, a hapless amnesiac. This will not stop me; my sense of purpose is so strong that I would gladly sacrifice *everything* to succeed.

I sit in the time machine's seat and strap myself in. I set the timer to go off in ten minutes. Off to 1913, a reasonable choice, as he should be much easier to kill before he comes to power, and I don't think I could bring myself to kill a child. I inject myself with the memory-wiping serum — and then I wait. And I wait.

A blinding light. Where am I? Who am I?

→ NATURE.COM
Follow Futures:

✓ @NatureFutures

☑ go.nature.com/mtoodm

Standing on an open field next to a wide road, the wind ruffling my hair, I can't



remember who I am but I know I have a glorious purpose in life. It feels both frightening and exhilarating.

I hear a faint moaning sound from down the road and I move towards it. The road turns abruptly to the left to cross a thicket of trees. As I enter the wood, I see an overturned automobile. One of the wheels lies a little farther down the road; it must have come off as the car turned the bend.

The moaning comes from inside the car. Filled with apprehension, I open the door to see a young man lying in a pool of blood. There's an ugly wound in his left temple.

"Water, please," he whispers when he sees me.

"I'm sorry, I have none," I tell him. "Try to hang on, I'll go for help."

He clasps my hand. "I'm going to Munich," he says. He coughs blood. "I'm a painter."

He's struggling to speak; then his eyes roll and he dies.

I rummage through his sack and find his papers. His name was Adolf Hitler.

I say it loudly. It has a nice ring to it; it even sounds vaguely familiar.

There's money in his wallet. I put it in my pocket. On a whim, I also take his papers. I lacked a name and now I have one. A name and a sense of purpose. I have a feeling that people will talk about me in years to come.

As I say the name again, weird, unnatural, terrible visions come unbidden to my mind. Is this to be my destiny? I shiver with fear — or anticipation.

George Nikolopoulos is a speculative fiction writer from Greece. His stories have been published in Galaxy's Edge, Factor Four, Best Vegan SFF 2016 and elsewhere.